Running head: Using MR equations built from summary data

Using regression equations built from summary data in the psychological assessment of the individual case: Extension to multiple regression

John R. Crawford
University of Aberdeen

Paul H. Garthwaite
Department of Mathematics and Statistics
The Open University

Annie K. Denham
University of Aberdeen

Gordon J. Chelune
Department of Neurology
University of Utah

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 psychological assessment. Moreover 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 psychologists are aware that regression equations can be built not only from raw data but also using only basic summary data for a sample, and (b) the computations involved are tedious and prone to error. In an attempt to overcome these barriers, Crawford and Garthwaite (2007) provided methods to build and apply simple linear regression models using summary statistics as data. In the present study we extend this work to set out the steps required to build *multiple* regression models 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 a computer program that implements these methods. Although there are caveats associated with the use of the methods, these need to be balanced against pragmatic considerations and against the alternative of either entirely ignoring a pertinent dataset or using it informally to provide a clinical "guesstimate". Upgraded versions of earlier programs for regression in the single case are also provided; these add the point and interval estimates of effect size developed in the present paper.

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

INTRODUCTION

Within today's health care environment, the term "evidence-based practice" has become common place, and was formally introduced in medicine in 1992 (Evidence-Based Medicine Working Group, 1992). At the heart of evidence-based practice is outcomes accountability guided by empirical research evidence within the context of clinical expertise and patient values (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). In 2006, the American Psychological Association (APA Presidential Task Force on Evidence-Based Practice, 2006) adopted a similar position stating that "clinical decisions (should) be made in collaboration with the patient, based on the best clinically relevant evidence, and with consideration for the probable costs, benefits, and available resource options" (p. 285).

The proposition that decision making in clinical practice should be based on objective, empirical data is not new. Paul Meehl's book, *Clinical Versus Statistical Prediction* (Meehl, 1954), not only identified clinical and actuarial approaches to data collection and individual prediction as distinctly different processes, but laid the basis for decades of comparative research that repeatedly demonstrates that: a) virtually any diagnostic question or prediction of behavior can be addressed by actuarial predictions, and b) empirically-based decision algorithms are almost always vastly superior to clinically-based decision making while being more reliable, accurate, and cost-effective (Dawes, Faust, & Meehl, 1989; Grove & Lloyd, 2006; Salzinger, 2005). Unfortunately, despite the strength of evidence favoring statistically-based actuarial methods, they have had only modest impact on everyday decision making (Dawes et al., 1989; Hamilton, 2001).

Chelune (2002 ; 2010) has argued that widespread adoption of evidence-based practices in clinical psychology would be facilitated if researchers and clinicians alike would embrace the tenets that: a) clinical outcomes are individual events that are can be characterized by a change in status, performance, or other objectively defined endpoint, and b) outcomes research must be analyzed and packaged in a manner that can be directly evaluated and applied by the clinician in the individual case. Too often, outcomes research in psychology as been limited to methods of null hypothesis significance testing that report only aggregate data on group differences, which are difficult for even informed clinicians to apply in the individual case. Fortunately, as recently reviewed by McIntosh and Brooks (2011), there are a growing number of statistical procedures for comparing the results of an individual patient against control samples, including procedures for constructing bivariate prediction equations derived from sample summary data in published research studies and test manuals and testing whether an individual's observed score is meaningfully different from his/her predicted score (Crawford & Garthwaite, 2006; 2007). The purpose of this paper is to expand this work to multiple regression-based predictions and to provide illustrative applications of the methods.

The roles for regression equations in the assessment of the individual case

Regression equations serve a number of useful functions in the psychological assessment of individual cases (Chelune, 2003; Crawford & Garthwaite, 2007; Crawford & Howell, 1998; Strauss, Sherman, & Spreen, 2006; Temkin, Heaton, Grant, & Dikmen, 1999). For example, regression equations are widely used to estimate premorbid levels of ability in clinical populations using psychological tests that are relatively resistant to psychiatric or neurological dysfunction (Crawford, 2004; Franzen, Burgess, & Smith-Seemiller, 1997; O'Carroll, 1995).

Regression is also commonly used in the assessment of change in cognitive functioning in the individual case (Crawford & Garthwaite, 2007). Here a regression equation is built (usually using healthy participants) to predict a case's retest score on a cognitive ability measure from their score at initial testing. A predicted retest score that is markedly higher than the obtained retest score suggests cognitive deterioration (Crawford & Garthwaite, 2007; Heaton & Marcotte, 2000; Sherman et al., 2003; Temkin et al., 1999).

Clinical samples can also 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 patients with intractable temporal lobe epilepsy who had not undergone any surgical intervention in the intervening period. The equation was then used to assess the effects of temporal lobectomy on memory functioning in a sample of surgical patients.

As Crawford and Garthwaite (2007) observed, "regardless of whether an equation is 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)" (p. 611).

Regression equations can also provide an alternative to the use of conventional stratified normative data (Heaton & Marcotte, 2000). For example, if performance on a neuropsychological test is affected by age and years of education (as is commonly the case), then these variables can be incorporated into a regression equation to obtain an individual's predicted score on the test. This use of regression provides what Zachary and Gorsuch (1985) have termed "continuous norms". Such norms can be contrasted with the discrete norms formed by creating arbitrary age by education bands. With the latter approach, a case's apparent relative standing can change dramatically as he/she moves from one age or education band to another (Crawford & Garthwaite, 2007).

It can be seen from the foregoing that regression equations perform many useful roles in the neuropsychological assessment of individuals. However, the potential of regression equations is far from being fully realized. As Crawford & Garthwaite (2007) note, "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 psychological functioning of individual cases" (p. 611). For example, there are literally hundreds of published studies that have examined performance at test and retest on a wide variety of commonly used psychological tests (see McCaffrey, Duff & Westervelt, 2000 for examples).

To enable psychologists to use published data in the assessment of the individual case Crawford and Garthwaite (2007) developed methods to build simple regression models (i.e., models that use a single predictor variable) from summary data: the resultant regression equation together with its associated statistics (such as the standard error of estimate which can also be calculated from summary data) can then be applied to specific cases to infer whether they exhibit a large and / or statistically significant difference between their obtained scores on a task and the score predicted by the equation. These authors implemented the methods in a computer program that takes summary data from a sample and an individual case's data as input, builds the equation, and then reports the results for the specific case.

Building *multiple* regression equations from summary data

Compared to regression equations with a single predictor, multiple regression equations provide a more flexible and potentially more sensitive means of testing hypotheses concerning an individual case. For example, when testing for change in an individual's cognitive functioning, if age or years of education are related to the magnitude of practice effects, then these variables can be incorporated into an equation along with the initial test score to obtain a more precise estimate of an individual's expected score at retest; this estimate can then be compared to the score actually obtained by the case (Duff et al., 2005; Temkin et al., 1999).

Crawford and Garthwaite (2006) provided inferential methods for comparing a case's obtained score with a predicted score from a multiple regression. However, these methods assume that the multiple regression equation and its associated statistics are available. That is, the methods take the intercept (a) for the equation plus the vector of beta values (**b**) and the equation's standard error of estimate ($s_{Y \cdot \mathbf{x}}$) as inputs.

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

Moreover, those psychologists who know that summary statistics are sufficient also know that the calculations involved are complicated, very time-consuming, and prone to error. Currently, therefore, in situations where multiple regression equations would be helpful and could be built, the vast majority of psychologists do not avail themselves of the opportunity. Alternatively, if a valiant psychologist does attempt to build an equation, 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 necessary methods deals with all of these problems.

The remainder of this paper has three principal aims. The first is to set out the calculations required to build multiple regression equations from summary data and outline the further calculations required when applying these equations to draw inferences concerning an individual case. The second aim is to describe and make available a computer program that implements all the methods described. The third aim is to provide examples of how these methods can be applied in psychological assessment.

We acknowledge that there will be fewer opportunities for psychologists to employ the current methods than those developed by Crawford and Garthwaite (2007) for simple linear regression. The limiting factor is that the reports providing the summary data need to contain not only the correlations of predictor variables with the criterion, which will be common, but also the correlation(s) *between* the predictor variables, which will be less common. The means and standard deviations for all variables are also required but these data are typically available in most research reports.

Method

Building a multiple regression equation from summary data

The multiple regression equation relating $\mathbf{x}=\left(X_1,X_2,\ldots,X_k\right)'$, the $k\times1$ vector of predictor variables, to Y, the criterion variable is

$$\begin{aligned} Y &= a+b_1X_1+b_2X_2+\ldots b_kX_k+\varepsilon \\ &= a+\mathbf{b}'\mathbf{x}+\varepsilon \qquad (1) \end{aligned}$$

where ε is the random error and $\mathbf{b}$ is a $k\times1$ vector. Assuming normality,

$$Y \sim N(a + \mathbf{b}'\mathbf{x},\ \sigma^2).$$

We want to obtain least-squares estimates, $\hat{a}$ and $\hat{\mathbf{b}}$, of a and $\mathbf{b}$, and $s^2_{Y \cdot \mathbf{x}}$ (the square of the standard error of estimate, i.e., the variance of the errors of estimate) as the estimate of σ^2, from summary data for a sample, namely the $(k+1)\times 1$ vector of means, and the $(k+1)\times 1$ vector of standard deviations for the *k* predictor (*X*) variables and *Y*, and the matrix of correlations. The first step is to partition the matrix into a $k \times k$ matrix of correlations, **R**, for the *X* variables, and a $1 \times k$ row vector of correlations of each *X* variable with *Y*, which we denote **r**. Also form a vector of the means for the *X* variables, $\bar{\mathbf{x}}$, and a vector of standard deviations for the *X* variables, **s**. Next invert **R** and post-multiply it by **r** to obtain the vector of standardized betas, $\hat{\mathbf{b}}_s$. That is,

$$\hat{\mathbf{b}}_s = \mathbf{R}^{-1}\mathbf{r}.$$

($\hat{\mathbf{b}}_s$ would be the vector of regression coefficients if the *X* variables and *Y* were standardized.) Next, divide **s** by the scalar quantity s_Y, (the standard deviation of the *Y* variable), that is compute $s_Y^{-1}\mathbf{s}$. Form a diagonal matrix, **S**, with the $s_Y^{-1}\mathbf{s}$ as the diagonal entries (all off-diagonal entries are zero in a diagonal matrix). By pre-multiplying $\hat{\mathbf{b}}_s$ by $\mathbf{S}^{-1}$ we obtain the $k \times 1$ vector of unstandardized betas, $\hat{\mathbf{b}}$. That is

$$\hat{\mathbf{b}} = \mathbf{S}^{-1}\hat{\mathbf{b}}_s. \qquad (2)$$

Also

$$\hat{\alpha} = \bar{Y} - \hat{\mathbf{b}}\bar{\mathbf{x}}. \qquad (3)$$

We now have the regression equation for predicting *Y* from the *X* variables obtained entirely from summary data.

To obtain the standard error of estimate ($s_{Y\cdot\mathbf{x}}$) for this equation we first obtain R^2 (the proportion of variance in *Y* explained by the *X*s). That is

$$R^2 = \hat{\mathbf{b}}_s\mathbf{r}.$$

Then

$$s_{Y\cdot\mathbf{x}} = \sqrt{\frac{\left(1-R^2\right)\left(s_Y^2\left[n-1\right]\right)}{n-k-1}}. \tag{4}$$

Supplementary statistics, such as the squared semi-partial correlations for each predictor variable, adjusted (shrunken) R^2 , and a test on the overall significance of the regression, are all obtained using the standard formulas so are not covered here; see Cohen, Cohen, West and Aiken (2003) or Tabachnick and Fidell (2005) for details. For present purposes it is sufficient to note that all these statistics can be obtained from summary statistics.

Inferential method for the discrepancy between a case's obtained and predicted scores

Having set out the steps to obtain a multiple regression equation from summary data we now turn to the calculations required to draw inferences concerning the discrepancies between a given case's obtained score on Y_O and the score predicted by such an equation, $\hat{Y}$. The following methods are those developed by Crawford and Garthwaite (2006) but are set out here for the convenience of the reader.

The first step is to calculate the standard error of a predicted score for a new case, which we denote as s_{n+1} (Crawford & Howell, 1998; Howell, 2002). This standard error can be expressed in a number of different but equivalent forms (Cohen et al., 2003); here we use the form set out in Crawford and Garthwaite (2006):

$$s_{n+1} = s_{Y\cdot\mathbf{x}}\sqrt{1+\frac{1}{n}+\frac{1}{n-1}\sum r^{ii}z_{io}^{2}+\frac{2}{n-1}\sum r^{ij}z_{io}z_{jo}}, \tag{5}$$

where r^{ij} identifies off-diagonal elements of the inverted correlation matrix ($\mathbf{R}^{-1}$) for the *k* predictor variables, r^{ii} identifies elements in the main diagonal, and $z_0 = \left(z_{10},\ldots,z_{k0}\right)'$ identifies the case's values on the predictor variables in *z* score form. The first summation in equation (5) is over the k diagonal elements and the second is over the $k\left(k-1\right)/2$ off-diagonal elements below (or above) the diagonal. Crucially for present purposes, it can be seen that this statistic can be calculated when only summary data from a sample are available.

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 **b** from a sample. It can be seen from equation (5) that s_{n+1} will increase in magnitude the further the case's scores on the predictor variables are from their respective means, as the components of z_0 will increase in magnitude; this is also a consequence of the uncertainty in estimating the betas.

Next one computes the discrepancy between a case's obtained score, Y_O, and their score predicted by the regression equation, $\hat{Y}$, and divides this discrepancy by s_{n+1} to yield a standardized discrepancy between the obtained and predicted score. That is

$$\frac{Y_O-\hat{Y}}{s_{n+1}}. \tag{6}$$

Under the null hypothesis, that the discrepancy is an observation from the population sampled to build the equation, the resultant quantity will have a *t*-distribution on $n-k-1$ df (Crawford & Garthwaite, 2006). 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.

Significance tests have a role to play in psychological assessment of the single case. When a discrepancy achieves statistical significance the psychologist can be confident that it is unlikely to be a chance finding; i.e., 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 are largely arbitrary conventions; the conclusion drawn when a case'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 psychologist should be primarily concerned with the more general issues of the degree of rarity of the case's discrepancy and, relatedly, with the *size* of the effect. In the remainder of this section we deal with the rarity of the discrepancy; the effect size issue is dealt with in the next section.

Fortunately, an estimate of the rarity of the case's discrepancy is an inherent feature of the method: the *p* value used to test significance is also a point estimate of the proportion of the relevant sub-population that would obtain a discrepancy more extreme than that observed for the case (Crawford & Garthwaite, 2006), where the relevant sub-population is the set of people with the same value on the predictor variable (i.e., X) as the case. As noted, the full 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 patients with a severe TBI six months post injury.

For a formal proof that the *p*-value from the significance test also equals the estimated proportion of the population exhibiting a more extreme discrepancy than the case see Appendix 1 of Crawford and Garthwaite (2007). When quantifying the rarity of a case's data 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 rarity 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 rarity of the discrepancy between an individual's obtained and predicted score. Crawford and Garthwaite (2006) 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 case.

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. However, 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, even when the predicted score is obtained from a multiple regression equation, they can be calculated without requiring the sample's raw data.

Confidence limits on the rarity 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.

Point and interval estimates of the effect size for the discrepancy between observed and predicted scores

A number of authorities in statistics and psychology have made strenuous calls for the reporting of indexes of effect size. For example, in a report on statistical inference, the American Psychological Association strongly endorsed the reporting of effect sizes. The report recommends that researchers should "always provide some effect-size estimate when reporting a *p*-value" and goes on to note that "reporting and interpreting effect sizes… is essential to good research" (Wilkinson and The APA Task Force on Statistical Inference, 1999, p. 599).

Advice aimed specifically at neuropsychologists has also been offered (e.g., Bezeau & Graves, 2001; e.g., Crawford & Henry, 2004; Zakzanis, 2001) and editorial policies requiring the reporting of effect sizes in psychology journals (Becker, Knowlton, & Anderson, 2005) have provided a further impetus. Although it is true to say that the take-up of such advice has been relatively slow, reporting of effect sizes in group-based psychological research is now fairly common (Crawford, Garthwaite, & Porter, 2010a).

Crawford and Garthwaite (2006) provided an index of effect size for the discrepancy between observed and predicted scores. However, this consisted of only a *point* estimate of effect size. In the present study we provide a slightly different standardized effect size index, which we denote as z_{OP}, and we accompany the point estimate with an *interval* estimate. To obtain the point estimate of the effect size put

$$z_{OP} = \frac{Y_O - \hat{Y}}{s_Y \sqrt{1 - R^2}}, \qquad (7)$$

where all terms have been defined previously. The *OP* subscript for this *z* value denotes that it an effect size for the discrepancy between a case's **O**bserved and **P**redicted scores, and is used to differentiate it from other effect size indexes developed for use with the single-case in Crawford et al. (2010a) and elsewhere (Crawford, Garthwaite, & Wood, 2010b). It can be seen from equation (7) that if z_{OP} is positive the case's obtained score exceeds the predicted score; if it is negative then the obtained score is lower than the predicted score.

The denominator in equation (7) will be familiar to many readers. It is the formula often used to represent the standard error of estimate but it is independent of sample size. This means that it is unsuitable for significance testing and other inferential purposes, where the full version of the standard error of estimate should be employed, i.e., equation (4) in the present paper. However, this is the very feature required for an index of effect size.

This effect size estimate is analogous to the use of *z* when comparing a case's score on a psychological test to that of a control or normative sample. That is, *z* tells us how many SDs units the case's score is above or below the normative mean. In the present case we can think of the discrepancy between the obtained and predicted score as a derived score. The mean discrepancy score in the sample used to build the equation is necessarily zero and z_{OP} tells us how many SDs the case's discrepancy is from this mean.

In group-based research there is an increasing recognition that *point* estimates of effect size should be accompanied by *interval* estimates (i.e., confidence intervals or credible intervals); e.g., see Steiger (2004), Fidler and Thompson (2001), and

Thompson (2007). That is, all statistics have uncertainties attached to them and effect sizes are no exception; these uncertainties should therefore be quantified when possible.

Crawford, Garthwaite and Porter (2010a) have argued that, in keeping with the general principle that the standards of reporting when working with individual cases should be as high as those expected in group-based research, interval estimates for effect sizes should also be reported for individual cases. Fortunately, for the present problem, the statistical theory necessary to form such interval estimates already exists. An intermediate step in Crawford and Garthwaite's (2006) method for setting 95% confidence limits on the rarity of a case's discrepancy (see previous section) involves generating two standard normal deviates, and these provide the required upper and lower 95% limits on the effect size index. The derivation of these limits on an effect size and the calculations required to obtain them are set out in Appendix 1 of the present paper. Monte Carlo simulations were conducted to verify that these confidence limits performed as they should; i.e., that they captured the true effect size on 95% of Monte Carlo trials (details of these results are available from the first author on request).

Results and Discussion

Implementing the methods in a computer program

A computer program for PCs was written to accompany this paper, and it implements all of the methods covered in the present paper. The program (RegBuild_MR.exe) prompts the user for the sample means and standard deviations of the criterion variable and predictor variables, the correlation matrix for these variables, and *n* for the sample. A screen capture of the data entry form is presented in Figure 1a; the data entered are those used in the first worked example (see later section).

The output is divided into two sections. The first records the results from performing the multiple regression, i.e., the unstandardized regression coefficients (**b**) and the intercept (*a*) for the regression equation, together with its standard error of estimate. The squared semi-partial correlation coefficient for each predictor variable is also recorded (to allow users to assess the *unique* contribution of each variable). The program also reports Multiple R, R^2, adjusted (shrunken) R^2, and the F value used to test for the significance of the regression with its accompanying p-value.

These outputs are followed by the results obtained from analyzing the individual case's data. These consist of: (a) the case's predicted score; (b) the discrepancy between the case's obtained and predicted scores; (c) the point and interval estimates of the effect size for the discrepancy (by default the 95% confidence limits on this percentage are two-sided, alternatively a one-sided upper or lower 95% limit can be requested); (d) the results of the significance test (one- and two-tailed probabilities are provided); and (e) the point estimate of the percentage of the population that would obtain a larger discrepancy with a confidence interval for this percentage (by default the 95% confidence limits are two-sided, 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 single case); these notes are reproduced in the output from the program. A screen capture of the output form for the computer program is presented in Figure 1b; the results are again those obtained for the first worked example (see later section). Note that not all of the results can be reproduced in a single screen capture: in the present case the beta values for the predictor variables are not shown.

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 case, 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.

A compiled version of this program can be downloaded (as an executable file or as a zip file of the executable) from the following website address: http://www.abdn.ac.uk/~psy086/dept/RegBuild_MR.htm.

Although one of the main aims of the methods set out in the present paper was to allow psychologists to build and use regression equations from summary data, a reviewer of an earlier version of this manuscript pointed out that it would also be useful to build and apply equations using *raw* data from a normative or control sample as inputs. We agree and have therefore written a companion program, Regbuild_MR_Raw.exe, to provide this capability. The raw data are read from a text file prepared by the user, in which the first *n* rows consists of the scores of the sample on the criterion variable and predictor(s), and the last (i.e., *n* + 1th) row consists of the corresponding scores for the case; full instructions on preparing this data file are provided in the program's information panel.

Upgrading earlier regression methods for the single case to incorporate point and interval estimates of effect size

Crawford and Garthwaite's (2007) methods and accompanying computer program (RegBuild.exe) for building and using regression equations for *bivariate* problems did not offer interval estimates of effect size for the discrepancy between obtained and predicted scores. Given the increasing emphasis placed on the use of both effect sizes and confidence intervals, we have upgraded the program to provide these point and interval estimates (and added an ES suffix to the program name, so it can be differentiated from the earlier version).

Crawford and Garthwaite (2007) also made a companion program available for the bivariate case that allowed regression to be performed even when the correlation between the predictor and criterion variable was unavailable, provided that results of a paired *t*-test or ANOVA comparing the predictor and criterion were reported. This program has also been upgraded to incorporate the point and interval estimates of effect size and has been renamed (Regbuild_t_ES.exe). Both of these upgraded programs can be downloaded from the same URL provided earlier.

Examples of the use of the methods and accompanying programs

In this section we illustrate some ways in which the methods and accompanying computer program can harness summary data from published studies in order to assist psychologists to draw inferences concerning the cognitive status of individual cases. In doing this we adopt the general examples used by Crawford and Garthwaite (2007) to illustrate the use of simple regression but extend these to include multiple predictor variables.

Suppose that a psychologist has seen a 60 year old male patient with 16 years of education because of suspected early Alzheimer’s disease. Further suppose that a semantic (category) fluency test had been administered at the initial assessment and again after five months and that an initial letter fluency test (e.g., FAS) was also administered at the first assessment. The case’s score on the semantic fluency test at initial testing was 28 and the score at retest was 24; the case’s FAS score at initial testing was 34.

~Tables 1 and 2 about here~

Table 1 sets out details of four hypothetical studies: for each study it lists the summary data required to build a multiple regression equation and to calculate the associated statistics for drawing inferences concerning an individual case. For reasons of space the correlations among the predictor variables and the criterion variable are reported separately in Table 2. 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 program is designed to be intuitive, the provision of the summary data in Tables 1 and 2 and the worked examples below will allow users to run these examples themselves. This will help users become familiar with the mechanics of the process prior to using the methods with their own data.

Study A was a study conducted on a sample of healthy participants (age range 50 to 80) on the effects of ageing on psychological test performance; in the course of this study the correlation between age and performance on the semantic fluency (SF) test (-0.56) was reported, as was the correlation between SF and years of education (0.66). It can be seen (Table 2) that both age and education exert a substantial effect on performance on the semantic fluency task.

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

Applying the methods set out earlier to build a multiple regression equation from the sample summary statistics yields the unstandardized regression coefficients and the intercept; these are reported in Table 1 (as noted, associated statistics for the multiple regression are also provided by the computer program that accompanies this paper; see the screen capture, Figure 1b, for these statistics for this particular example). Applying the regression equation to the case, his *predicted* semantic fluency score, based on his age and years of education is 52.15. Using equations (4) and (5), the standard error of estimate for this equation ($s_{Y\cdot\mathbf{x}}$) is 7.433 and the standard error for an additional individual (s_{n+1}) is 7.487 (these statistics are reported in Table 1, as are the equivalent statistics for the subsequent worked examples). The difference between these two statistics is modest in this example because the case's values on the predictor variables (i.e., his age and years of education) are not very extreme relative to the sample means 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 case's obtained semantic fluency score of 28 and predicted score of 52.15 is −24.15. Dividing this discrepancy by s_{n+1} yields a value of −3.226. Under the null hypothesis this difference is distributed as *t* on $n-k-1=180-2-1=177$ 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 = 0.0007, one-tailed).

The point estimate of the rarity of this discrepancy (i.e., the percentage of the population that would be expected to exhibit a discrepancy larger than that observed) is 0.075%. The accompanying 95% confidence interval on the percentage of the population that would exhibit a larger discrepancy than the patient ranges from 0.013% to 0.23%. Thus, in summary: there is a very large and significant discrepancy between the case's predicted and obtained scores. This size of discrepancy is estimated to be very unusual in the healthy elderly population and is consistent with severely impaired performance on the semantic fluency task.

Finally, before leaving Study A, the effect size for the discrepancy between the obtained and predicted score is very large $z_{OP} = -3.268$ (95% CI = −3.660 to −2.836). If, rather than using regression to compare the case's obtained and predicted scores, the case was simply compared to the mean of the sample in Study A, the case's performance would not look nearly as extreme. The effect size for such a comparison is $z = -1.17$; thus, although the case is just over one SD below the "normative" mean of the sample, this difference is modest compared to that obtained when the regression equation is used to provide an individualized comparison standard.

In this example, the use of regression served to expose a severe impairment. However, it will be appreciated that the use of regression may also help avoid incorrectly inferring the presence of an acquired impairment. For example, for the present data, the performance of a case who obtains a low score may not look very unusual if they were substantially older and had a modest number of years of education.

Moving on to Study B: this study was also a study of cognitive ability in healthy elderly participants and included among its results the correlation between the semantic fluency test and the FAS test, as well as the correlation of both tests with years of education (see Table 2 for the correlations and Table 1 for the other summary data for this second study). In psychological 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).

As Crawford and Garthwaite (2007) note, 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). Therefore, if an individual exhibits a large discrepancy between these two tasks, this suggests an acquired impairment on the more poorly performed task.

In this example there is an additional, specific, consideration: there is good evidence that semantic fluency performance is more severely disrupted by Alzheimer's disease (AD) than is initial letter fluency. For example, a 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. Rascovsky, Salmon, Hansen, Thal and Galasko (2007) also demonstrated the clinical utility of discrepancies between semantic and letter fluency in differentiating patients with autopsy-confirmed cases of Alzheimer's disease and frontotemporal lobar degeneration. 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 years of education, and then compare the individual's predicted and obtained scores. The regression equation and associated statistics built with the hypothetical data from Study B are presented in Table 1. Based on his initial letter fluency score of 34 and his 16 years of education, the case's predicted semantic fluency score is 45.74 using this equation, which is substantially higher than his observed score of 28.

Dividing the raw discrepancy between the obtained score and predicted score (-17.74) by s_{n+1} gives a value of –2.33. Evaluating this *t*-value reveals that the case's obtained score is significantly below the score predicted from his initial letter fluency score (*p* = 0.0108, one-tailed). The point estimate of the rarity of this discrepancy (i.e., the percentage of the healthy elderly population that would be expected to exhibit a discrepancy larger than that observed) is thus 1.08% and the 95% confidence interval is from 0.28% to 2.7%.

In summary, the case's semantic fluency is considerably lower than expected given his years of education and initial letter fluency performance; the discrepancy is very unusual and is consistent with a marked differential deficit in semantic versus initial letter fluency. As was the case in the first worked example, the effect size for the discrepancy between obtained and predicted scores, z_{OP} = −2.38 (95% CI = −2.77 to −1.93), is large. Again, this effect is much larger than would be obtained if the case was simply compared to the sample mean for semantic fluency in Study B (*z* = −1.27).

Note that a 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 (where an *a priori* decision to use a one-tailed test was made) 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 0.022.

Turning to Study C, psychologists commonly have to attempt to detect change in cognitive functioning in the individual case, for example, to monitor the positive or negative effects of interventions, to determine whether there is recovery following a stroke or TBI, or to detect decline in degenerative conditions. Serial assessment plays a particularly important role in the diagnosis of AD, 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: A psychologist need only find test-retest data for the measures used in an appropriate sample retested over an interval similar to that of their patient. Although regression can be used to predict scores at retest solely from initial test scores, it has quite commonly been found that other variables (normally demographic variables, such as age or years of education) can explain variance in retest scores over and above that explained by initial scores (Duff et al., 2005; Temkin et al., 1999).

Study C is a hypothetical test-retest study in which a sample of healthy elderly participants (N = 70) were tested on the semantic fluency test and retested after 6 months (this test-retest interval is a slightly longer than the interval for the case but sufficiently close to justify use of the data). Table 1 presents the summary statistics for this sample; the correlation matrix for the Study is presented in Table 2. Table 1 also presents the resultant multiple regression equation together with its associated statistics.

It can be seen from Table 1 that, in the healthy elderly sample, there was a practice effect (the mean at retest was 47.2, compared to the mean at first testing of 43.2. Using the regression equation, the case's predicted semantic fluency score at retest is 39.8, based on his age and initial score of 28. The score is below the mean score at retest because the case's initial test score was low. However, the predicted score at retest is still well above the case's obtained score at retest of 24. Dividing the raw discrepancy between the obtained score and predicted score (-15.8) by s_{n+1} yields a value of –1.92. Evaluating this *t*-value reveals that the patient's retest score is significantly below the score predicted from his score on first testing ($p = 0.0297$, one-tailed).

The point estimate of the rarity of this discrepancy is thus 2.97% and the 95% confidence limits on the percentage are from 0.48% to 8.7%. 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 D 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. In this study summary data for years of education was available and education was found to be a predictor of retest scores. Having obtained evidence of a decline for the case using the equation built using data from Study C, the data from Study D are used to examine whether or not the change from test to retest is unusual for patients with AD.

Using the data from Study D to build a regression equation, the patient's predicted semantic fluency score at retest is 23.6 (based on his initial score of 28 and 16 years of education) compared to his obtained retest score of 24. It is immediately clear that, although the case's initial test score and retest score are both higher than the corresponding AD sample means, the discrepancy between the obtained and predicted scores is very minimal. In this case it is not necessary to formally analyze the data but for completeness, dividing the raw discrepancy between the obtained score and predicted score (-0.4) by the s_{n+1} yields a standardized difference of – 0.04. Evaluating this *t*-value reveals that the patient's obtained retest score is clearly not significantly different from the predicted score on first testing ($p = 0.987$, two-tailed). Thus, although from the analysis of the data from the preceding study, the patient has shown evidence of decline, the decline is very typical of AD.

In this example the discrepancy does not even approach significance on a two- or one-tailed test. In cases where the discrepancy was more substantial it would be appropriate to use a two-tailed test. That is, even if a psychologist had independent grounds to believe that a case's cognitive decline would be atypically rapid for AD, or atypically slow, it is unlikely that she/he would have sufficient confidence in this to rule out the alternative possibility.

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 psychologist 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 D 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.

The methods should not be regarded as simply providing a test of the null hypothesis

When a discrepancy between an obtained and predicted scores is statistically significant the psychologist can be particularly confident that a problem has been uncovered, or, when the obtained score exceeds the predicted score, that a genuine improvement in performance has occurred. In these circumstances we can reject the null hypothesis that the discrepancy was an observation from the distribution of discrepancies in the population sampled to build the equation.

However, as noted earlier, we suggest that the principal focus with the current methods should be on the degree of rarity of the discrepancy and its effect size, rather than whether the *p* value falls below or above the cusp for conventional statistical significance. For example, suppose that in one of the foregoing examples using Study A, B, or C, the discrepancy between the case's obtained and predicted score did not achieve statistical significance ($p > 0.05$) but the discrepancy was still fairly unusual and the effect size substantial. This would still constitute useful evidence and should be given weight when arriving at a formulation for the case, particularly if the results are consistent with information obtained by other means (i.e., from other test results, behavioural observations, or the case history).

Effect size estimates for the discrepancy between predicted and obtained scores

The foregoing discussion has illustrated the use of the effect size estimates developed in the present study. In this section we briefly discuss some specific issues associated with the use of these estimates. The point estimate of the effect size expresses the discrepancy in standardized units which is a basic and, we hope, useful piece of information for clinicians. That is, it tells the user how many standard deviation units the case's discrepancy is from the average (= 0) discrepancy in the control or normative sample (because it is an effect size it, unlike the other statistics provided, the point estimate intentionally treats the control data as fixed).

These features of the effect size mean that it can be usefully employed to compare a case's results from other regression equations built using the same or different samples. Moreover, as the effect size is expressed in standard units, it also provides a means of comparing a case discrepancy with results from other testing. For example, if a regression equation has been used to provide an individualized norm for a case's score on a particular test (using, say, age, gender and education as predictors) then the effect size for the discrepancy between the predicted and obtained score can be compared with a case's standardized (*z*) scores on other tests that have been obtained using conventional normative data.

The verbal labels "small", "medium", and "large" have been used to classify effect sizes (e.g., Cohen's *d*) for *group* comparisons. We do not think it would be appropriate to attach verbal labels to the effect size index provided for the individual case in the present study because, as has been illustrated, the regression methods can be applied to very diverse assessment problems and so one size could never fit all. Note also that, although Jacob Cohen provided the foregoing verbal classification system for group comparisons, he was ambivalent about doing so (Cohen, 1988).

Caveats on the use of these methods and some pragmatic considerations

As is true when applying any regression equation (whether built from summary data or from raw data) psychologists should be aware that an equation should only be applied to individual cases if their values on the predictor variables lie within the range of values obtained in the sample used to build the equation. For example, if an equation was built in a sample of the elderly and uses age as a predictor, then it would clearly be inappropriate to use the equation to draw inferences concerning middle-aged or young cases.

It should also be noted that 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 badly violated (see Tabachnick & Fidell, 2005 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 variables 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 predictors 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.

A further practical issue is that, even when the correlation matrix is available, the precision with which the correlations are reported will be more of a concern in using the present multiple regression method than it is in the bivariate case (Sokal & Rohlf, 1995); this would be especially so if many predictor variables were employed.

However, as noted by Crawford and Garthwaite (2007), these concerns need to be balanced by pragmatic considerations. First, with many of the combinations of predictors and criterion variables likely to be employed in practice there is little evidence that heteroscedasticity is a pervasive problem. For example, if the predictors and criterion variables are 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. Moreover, it should be remembered that, although the conditional distribution of the criterion variable is assumed to be normal, the predictor variables in regression problems can have any distribution: i.e., they do not need to be normally distributed and indeed can be simple dichotomies such as male (coded as, say, 0) versus female (coded as, say, 1).

Second, with regard to the possibility of non-linear relationships between the criterion variable and the predictors: this is perhaps most likely to be an issue when age is used as a predictor of cognitive test scores. However, although the strategy of incorporating any non-linear component (by using polynomial functions of age) into the equations is not available when summary data are used as inputs, it is likely that most of the relationship will be approximately linear. Thus, although in these circumstances a more accurate prediction of scores could be achieved, incorporating the linear component will still be a considerable improvement over ignoring age effects entirely. In circumstances where the raw data for the control or normative sample *are* available, then it is possible, by using the companion program described earlier (RegBuild_MR_Raw.exe), to incorporate polynomial functions of the predictors as additional data columns (e.g., if age is one of the predictor variables, then age^2 and even age^3 could also be entered as additional predictors).

Third, with regard to the precision with which the equation can be estimated, we envisage that typically the number of predictors would be relatively modest (two or three) in most applications of the current methods so that this is not as serious an issue as it might be. If a psychologist is concerned with this issue it would be relatively easy to check whether precision is an issue. Because the computer program accompanying this paper can be used very rapidly a user can easily re-run the analysis substituting the upper and/or lower range of the correlations. For example, suppose the correlations have been reported to two decimal places and that a correlation between a given predictor and the criterion was reported as 0.62, then this could be re-run substituting 0.625 or 0.615 and the effects quickly examined.

Fourth, and most importantly: in an ideal world, psychologists would routinely employ the principles of evidence-based practice and avail themselves of relevant research that would best inform their evaluations of individual patients (Chelune, 2010). They would also 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 published equations is very limited in comparison to: (a) the wide variety of hypotheses that psychologists may wish to test, and (b) the large reservoir of studies that report summary data on psychological tests.

Therefore, in the absence of an existing equation, and when relevant summary data are available, the evidence-based approach suggested here needs to be contrasted with the alternatives open to the psychologist. These are that the psychologist 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” (Crawford & Garthwaite, 2007). For example, in the latter case the reasoning might proceed along the following lines: “given that this test has a fairly high test-retest correlation, is subject to a moderate practice effect, and noting that age influences the magnitude of this effect and my case is relatively young, the difference between this case’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 differences expected by chance and we may overweight some variables at the expense of others.

Two forms of hypothesis test when examining discrepancies between predicted and obtained scores

The hypothesis test implemented in the present paper tests whether we can reject the null hypothesis that the discrepancy between predicted and obtained scores obtained by a case is an observation from discrepancies in the control population (as noted elsewhere the control population will most commonly be defined to be a healthy control population but need not be as is demonstrated in the final worked example using Study D).

There is however, an alternative form of null hypothesis test that can be applied to discrepancies between predicted and obtained scores: namely a test that the discrepancy is significantly different from zero. In other words, we could test whether any observed discrepancy between predicted and obtained scores is large enough for us to be confident that it does not simply reflect the effects of measurement error in the predictor and criterion variables. Reynolds (1984) provided equations for this latter form of hypothesis test.

This is a useful test and can be seen as a complimentary method than a competitor for the present form of hypothesis test. In essence the two forms of test address the two fundamental questions that arise when examining differences obtained for a case: are the differences *reliable* (the Reynolds test), and are the differences *abnormal* (such that we can reject the hypothesis that the case's discrepancy is an observation from the control population).

It will typically be the case that the Reynolds (1984) test will require smaller discrepancies to record a significant result. Indeed, if the variables involved have very high reliabilities, it will be common for individuals to exhibit significant (i.e., reliable) differences between their predicted and observed scores. It should be noted that the Reynolds (1984) method is a large sample method as, unlike the present methods, it assumes that the summary statistics for the variables (and their reliability coefficients) are fixed and known. It is therefore eminently suitable for use with standardized test batteries but would need to be used with caution if applied to data obtained from modestly sized samples. An excellent example of the application of this latter form of hypothesis test can be found in Schneider (2010a), where it is applied to scores on the Woodcock Johnson Tests of Cognitive Abilities – Third Edition (Woodcock, McGrew, & Mather, 2001); see Schneider (2010b) for further details.

Reporting of summary data in psychological studies

The emphasis in the foregoing sections has been on the application of the regression methods with summary data from existing studies. However, it is to be hoped that the availability of these methods will help encourage researchers to provide the full range of summary data when reporting their findings (either in the Results section, or as supplementary material).

For example, studies reporting on predictors of test performance in the general population, or predictors of outcome as measured by psychological tests in clinical populations clearly provide a useful knowledge base for the practicing psychologist. However, by including summary data, the utility of such studies data can be greatly enhanced as this would allow psychologists to directly apply the results to their individual cases. Reporting of summary data (e.g., reporting of the correlation matrix for studies using regression to examine group level effects) would also be in keeping with recommendations by the American Psychological Association (2001).

Applications in other areas of psychological practice

The examples used to illustrate the applications of the present methods have focused on clinical assessment issues. However, the methods are just as applicable to other areas of applied psychology in which assessments of the individual case are conducted. Obvious examples are industrial/ organizational/ occupational psychology and educational psychology. These areas have experienced just as large an increase in the amount of published data available to practitioners and therefore would hopefully also benefit equally from the opportunity to directly employ such data for inference at the level of the individual case.

References

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

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Beach, L. R., & Braun, G. P. (1994). Laboratory studies of subjective probability: A status report. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 107-128). Chichester, UK: Wiley.

Becker, J. T., Knowlton, B., & Anderson, V. (2005). Editorial. *Neuropsychology, 19*, 3-4.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research *Journal of Clinical and Experimental Neuropsychology, 23*, 399-406.

Chelune, G. J. (2002). Making neuropsychological outcomes research consumer friendly: A commentary on Keith et al. (2002). *Neuropsychology, 16*, 422-425.

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. (2010). Evidence-based research and practice in clinical neuropsychology *The Clinical Neuropsychologist, 24*, 454-467.

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.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioural sciences* (3rd ed.). Mahwah, NJ: Lawrence 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: 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. (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: a significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology, 20*, 259-271.

Crawford, J. R., & Garthwaite, P. H. (2007). Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology, 21*, 611-620.

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010a). Point and interval estimates of effect sizes in the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology, 27*, 245-260.

Crawford, J. R., Garthwaite, P. H., & Wood, L. T. (2010b). The case controls design in neuropsychology: Inferential methods for comparing two single cases. *Cognitive Neuropsychology, 27*, 377-400.

Crawford, J. R., & Henry, J. D. (2004). Assessment of executive deficits. In P. W.

Halligan & N. Wade (Eds.), *The effectiveness of rehabilitation for cognitive deficits* (pp. 233-245). London: Oxford University Press.

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.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Duff, K., Schoenberg, M. R., Patton, D., Paulsen, J. S., Bayless, J. D., Mold, J., et al. (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology, 20*, 281-290.

Evidence-Based Medicine Working Group. (1992). A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*, 2420-2425

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 575-604.

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

Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology, 115*, 192-194.

Hamilton, J. D. (2001). Do we under utilise actuarial judgement and decision analysis? *Evidence Based Mental Health, 4*, 102-103.

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.

McIntosh, R. D., & Brooks, J. L. (2011). Current tests and trends in single-case neuropsychology. *Cortex, 47,* 1151-1159.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

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. 301-318). Chichester: Wiley.

O'Carroll, R. (1995). The assessment of premorbid ability: A critical review.

Neurocase, 1, 83-89.

Rascovsky, K., Salmon, D. P., Hansen, L. A., Thal, L. J., & Galasko, D. (2007). Disparate letter and semantic category fluency deficits in autopsy-confirmed frontotemporal dementia and Alzheimer's disease *Neuropsychology, 21*, 20-30.

Reynolds, C. R. (1984). Critical measurement issues in learning disabilities. *Journal of Special Education, 18*, 451-477.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd ed.). New York: Churchill Livingston.

Salzinger, K. (2005). Clinical, statistical, and broken-leg predictions. *Behavior and Philosophy, 33*, 91-99.

Schneider, W. J. (2010a). *The Compositator 1.0*: WMF Press.

Schneider, W. J. (2010b). *The Compositator 1.0: User's Guide*: WMF Press.

Sherman, E. M. S., Slick, D. J., Connolly, M. B., Steinbok, P., Martin, R., Strauss, E., et al. (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.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*, 164-182.

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. (2005). *Using multivariate statistics* (5th ed.). New York: Pearson.

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.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.

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.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III.* Itasca: Riverside Publishing.

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

Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology, 16*, 653-667.

Appendix 1.

Derivation of the confidence limits on the effect size for discrepancies between obtained and predicted scores

The confidence intervals are based on theory developed by Crawford and Garthwaite (2006) and are derived from a non-central *t*-distribution. This distribution is defined by

$$T_{\nu}(\delta) = (Z+\delta)/\sqrt{\phi/\nu},$$

where *Z* has a normal distribution with a mean of zero and variance 1, and ϕ is independent of *Z* with a chi-square distribution on ν degrees of freedom. δ is referred to as the non-centrality parameter and effects the shape and skewness of the distribution.

We observe values $\mathbf{x}_0$ and Y_0^*. We require a $100(1-\alpha)\%$ confidence interval for z_{OP}^*, the true effect size, when a sample of size *n* gives estimates $\hat{a}$, $\hat{\mathbf{b}}$ and $s_{Y\cdot\mathbf{x}}^2$ for *a*, **b** and σ^2. Let $\hat{Y}_0$ be the predicted value of an individual whose **x**-values are $\mathbf{x}_0$, where the prediction is based on the sample used to build the equation. Put

$$\hat{Y}_0 = \hat{a} + \hat{\mathbf{b}}'\mathbf{x}_0, \tag{8}$$

where $\hat{a}$ and $\hat{\mathbf{b}}$ are the estimates of the regression coefficients in equations (3) and (4). Now put

$$\theta = \frac{1}{n} + \frac{\sum r^{ii} z_{io}^2 + 2\sum r^{ij} z_{io} z_{jo}}{n-1}, \tag{9}$$

where all right hand terms have previously been defined when presenting equation (5) in the main text. It is shown in Appendix 1 of Crawford and Garthwaite (2006) that

$$\operatorname{var}\left(\hat{Y}_0\right) = \sigma^2\theta.$$

Then we have we have that

$$\hat{Y}_0 \sim N\left(a + \mathbf{b}\mathbf{x}_0, \sigma^2\theta\right).$$

Let

$$z_{OP} = \frac{Y_0^* - \hat{Y}_0}{s_{Y.\mathbf{x}}}. \tag{10}$$

and put

$$z_{OP}^* = \frac{Y_0^* - a - \mathbf{b}\mathbf{x}_0}{\sigma}. \tag{11}$$

Then z_{OP} is an estimate of z_{OP}^*. Now,

$$\frac{z_{OP}}{\sqrt{\theta}} = \frac{\left(a + \mathbf{b}\mathbf{x}_0 - \hat{Y}_0\right)/\sqrt{\sigma^2\theta} + \left(Y_0^* - a - \mathbf{b}\mathbf{x}_0\right)/\sqrt{\sigma^2\theta}}{\sqrt{s_{Y\cdot\mathbf{x}}^2/\sigma^2}}, \tag{12}$$

and

$$\left(n - k - 1\right) s_{Y.\mathbf{x}}^2 / \sigma^2 \sim \chi_{n-k-1}^2.$$

Hence, $z_{OP}/\sqrt{\theta}$ has a non-central *t*-distribution with non-centrality parameter $\delta = z_{OP}^*/\sqrt{\theta}$ and $n - k - 1$ df. The $100\left(\alpha/2\right)\%$ and $100\left(1-\alpha/2\right)\%$ points of this distribution will depend on the value of δ. Let δ_L denote the value of δ for which the $100\left(1-\alpha/2\right)\%$ point is $z_{OP}/\sqrt{\theta}$. Similarly, let δ_U denote the value of δ for which the $100\left(\alpha/2\right)\%$ point is $z_{OP}/\sqrt{\theta}$. Then $(\delta_L\sqrt{\theta},\ \delta_U\sqrt{\theta})$ is a $100\left(1-\alpha\right)\%$ confidence interval for z_{OP}^*.

Table 1. Illustrative examples of the use of summary data from published studies to draw inferences concerning a case: summary data from four hypothetical studies are presented, together with the statistics for the resultant multiple regression equations (the correlation matrices for these data are presented in Table 2)

	Study A	Study B	Study C	Study D
SF Mean	41.3	43.4	47.2	20.2
SF SD	11.40	12.14	12.20	13.1
Predictor 1	Age	IF	SF Time 1	SF Time 1
Predictor 1 mean	66.8	36.6	43.2	24.3
Predictor 1 SD	8.42	12.50	11.20	12.10
Predictor 2	Education	Education	Age	Education
Predictor 2 mean	12.50	13.00	65.3	12.30
Predictor 2 SD	3.00	3.20	7.50	2.90
Sample size (*N*)	180	120	70	52
b_1	−0.539	0.540	0.652	0.691
b_2	2.055	1.226	−0.469	0.222
Intercept (α)	51.60	7.43	49.66	0.676
$s_{Y \cdot \mathbf{x}}$	7.433	7.53	7.96	10.02
s_{n+1}	7.487	7.61	8.23	10.28

Note: SF = semantic fluency; IF = initial letter fluency.

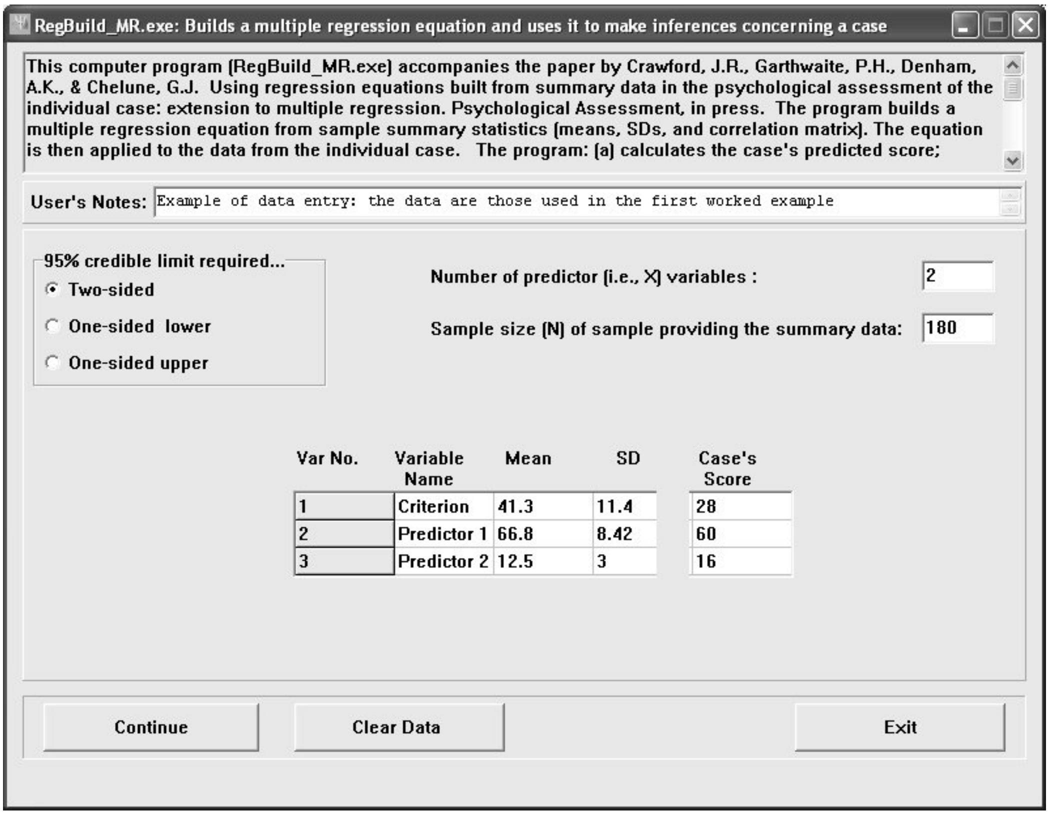
Table 2. Correlation matrices for the illustrative studies presented in Table 1

	Criterion (*Y*)	Predictor 1	Predictor 2
Study A			
Criterion (*Y*)	1.00		
Predictor 1	−0.56	1.00	
Predictor 2	0.66	−0.30	1.00
Study B			
Criterion (*Y*)	1.00		
Predictor 1	0.74	1.00	
Predictor 2	0.64	0.56	1.00
Study C			
Criterion (*Y*)	1.00		
Predictor 1	0.72	1.00	
Predictor 2	−0.54	−0.42	1.00
Study D			
Criterion (*Y*)	1.00		
Predictor 1	0.66	1.00	
Predictor 2	0.33	0.44	1.00

Figure Legends

Figure 1. Screen captures of (a) the input form, and (b) output form for the principal computer program that accompanies the present paper.

(a)



RegBuild_MR.exe: Builds a multiple regression equation and uses it to make inferences concerning a case

This computer program (RegBuild_MR.exe) accompanies the paper by Crawford, J.R., Garthwaite, P.H., Denham, A.K., & Chelune, G.J. Using regression equations built from summary data in the psychological assessment of the individual case: extension to multiple regression. Psychological Assessment, in press. The program builds a multiple regression equation from sample summary statistics (means, SDs, and correlation matrix). The equation is then applied to the data from the individual case. The program: (a) calculates the case's predicted score;

User's Notes: Example of data entry: the data are those used in the first worked example

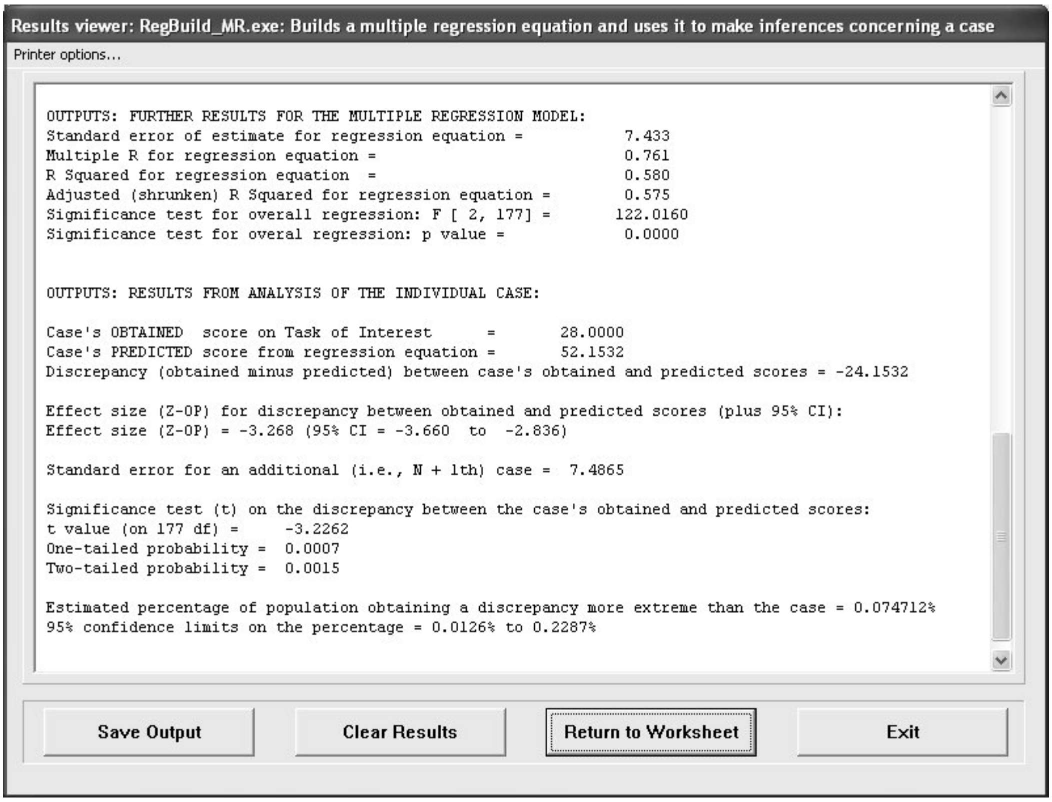
95% credible limit required...
- ● Two-sided
- ○ One-sided lower
- ○ One-sided upper

Number of predictor (i.e., X) variables : 2

Sample size (N) of sample providing the summary data: 180

Var No.	Variable Name	Mean	SD	Case's Score
1	Criterion	41.3	11.4	28
2	Predictor 1	66.8	8.42	60
3	Predictor 2	12.5	3	16

(b)



Results viewer: RegBuild_MR.exe: Builds a multiple regression equation and uses it to make inferences concerning a case

```
OUTPUTS: FURTHER RESULTS FOR THE MULTIPLE REGRESSION MODEL:
Standard error of estimate for regression equation =          7.433
Multiple R for regression equation =                          0.761
R Squared for regression equation  =                          0.580
Adjusted (shrunken) R Squared for regression equation =       0.575
Significance test for overall regression: F [ 2, 177] =     122.0160
Significance test for overal regression: p value =            0.0000


OUTPUTS: RESULTS FROM ANALYSIS OF THE INDIVIDUAL CASE:

Case's OBTAINED  score on Task of Interest      =       28.0000
Case's PREDICTED score from regression equation =       52.1532
Discrepancy (obtained minus predicted) between case's obtained and predicted scores = -24.1532

Effect size (Z-OP) for discrepancy between obtained and predicted scores (plus 95% CI):
Effect size (Z-OP) = -3.268 (95% CI = -3.660  to  -2.836)

Standard error for an additional (i.e., N + 1th) case =  7.4865

Significance test (t) on the discrepancy between the case's obtained and predicted scores:
t value (on 177 df) =      -3.2262
One-tailed probability =  0.0007
Two-tailed probability =  0.0015

Estimated percentage of population obtaining a discrepancy more extreme than the case = 0.074712%
95% confidence limits on the percentage = 0.0126% to 0.2287%
```