Neuropsychology
2006, Vol. 20, No. 3, 259–271

0894-4105/06/$12.00 DOI: 10.1037/0894-4105.20.3.259

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

John R. Crawford
University of Aberdeen

Paul H. Garthwaite
The Open University

In contrast to the standard use of regression, in which an individual's score on the dependent variable is unknown, neuropsychologists are often interested in comparing a predicted score with a known obtained score. Existing inferential methods use the standard error for a new case (s_{N+1}) to provide confidence limits on a predicted score and hence are tailored to the standard usage. However, s_{N+1} can be used to test whether the discrepancy between a patient's predicted and obtained scores was drawn from the distribution of discrepancies in a control population. This method simultaneously provides a point estimate of the percentage of the control population that would exhibit a larger discrepancy. A method for obtaining confidence limits on this percentage is also developed. These methods can be used with existing regression equations and are particularly useful when the sample used to generate a regression equation is modest in size. Monte Carlo simulations confirm the validity of the methods, and computer programs that implement them are described and made available.

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

Regression equations serve a number of useful functions in the neuropsychological assessment of individuals. Perhaps their most common role is as an alternative to the use of conventional normative data (Heaton & Marcotte, 2000). For example, if it is found that age, years of education, and gender are related to performance on a memory test, then a regression equation can be built (in a healthy sample) that uses these demographic variables as predictors. A patient's predicted score reflects her or his particular combination of demographic characteristics and thus provides an individual comparison standard for the score obtained on testing. Such an approach is in keeping with the emphasis placed on individual versus normative comparison standards in neuropsychological assessment (Crawford, 1996; Heaton, Grant, Ryan, & Matthews, 1996; Lezak, Howieson, Loring, Hannay, & Fischer, 2004). Even with a single predictor, such as age, a good case can be made for the regression approach over the use of conventional normative data. It provides what Zachary and Gorsuch (1985) have termed *continuous norms* rather than the discrete norms formed by creating arbitrary age bands; in the latter case, the relative standing of an individual can change dramatically as she or he moves from one age band to another.

Regression equations are also 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). Another common application of regression is in the assessment of change in cognitive 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 neuropsychological instrument at retest from her or his 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, Heaton, Grant, & Dikmen, 1999). Note that 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 persons with temporal lobe epilepsy 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).

A crucial commonality in all these examples is that they are concerned with examining the difference between the score predicted by the equation and the patient's actual score obtained on testing. That is, in all cases, the score on the dependent variable (or

John R. Crawford, School of Psychology, University of Aberdeen; and Paul H. Garthwaite, Department of Statistics, The Open University.

Additional materials are on the Web at http://dx.doi.org/10.1037/0894-4105.20.3.259.supp

We are grateful to Dr. Sytse Knypstra of the Department of Econometrics, University of Groningen, The Netherlands, for providing an algorithm that finds the noncentrality parameter of a noncentral *t* distribution given a quantile, its associated probability, and the degrees of freedom. This algorithm was used in Monte Carlo simulations and is also incorporated into the computer programs that implement the methods presented in this paper.

Correspondence concerning this article should be addressed to Professor John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, United Kingdom. E-mail: j.crawford@abdn.ac.uk

criterion variable, as it is also commonly termed) is known, and it is the discrepancy between this known obtained score and the predicted score that is of interest. This is very different from the standard or conventional use of a regression equation to predict an individual's standing on the dependent variable. In the standard application of regression, the score or value on the dependent variable is unknown. Such standard applications do arise in the assessment of individuals in other areas of psychology (e.g., regression may be used to predict later job performance from psychometric test scores), but they are relatively rare in neuropsychological settings. Unfortunately, existing inferential methods and discussion about the use of regression in statistics or biometrics textbooks are aimed almost exclusively at dealing with the standard applications. For example, the standard error of a predicted score for a new case (s_{N+1}), that is, a case not included in the sample used to build an equation, can be used to provide confidence limits on a predicted score (Cohen, Cohen, West, & Aiken, 2003; Crawford & Howell, 1998b; Howell, 2002) and hence is tailored to the standard usage. That is, these confidence limits are used to quantify the uncertainty associated with a predicted score on the criterion variable when this score is unknown.

The remainder of this paper is concerned with statistical methods for drawing inferences about the discrepancy between an individual's predicted and obtained score. Extending upon an observation made by Crawford and Howell (1998b), it will be shown that the s_{N+1} can readily be employed to provide a significance test on this discrepancy. Under the null hypothesis, the discrepancy observed for the patient is an observation from the population sampled to build the equation. In addition, the method described simultaneously provides a point estimate of the abnormality or rarity of a patient's discrepancy. That is, it estimates the percentage of the population that would obtain a more extreme discrepancy. This estimate should be very useful for neuropsychologists because their concerns go beyond simply whether the discrepancy is statistically significant.

It would also be useful to have a confidence interval to accompany the point estimate of abnormality. There is general agreement that confidence intervals should be used where possible (American Psychological Association, 2001; Daly, Hand, Jones, Lunn, & McConway, 1995; Gardner & Altman, 1989; Zar, 1996). For example, the American Psychological Association (2001) notes that "The use of confidence intervals. . .is strongly recommended" (p. 22). Similarly, Gardner and Altman (1989), in discussing the general issue of the error associated with sample estimates, note that "these quantities will be imprecise estimates of the values in the overall population, but fortunately the imprecision itself can be estimated and incorporated into the findings" (p. 3).

Neuropsychologists will be aware that estimates of the rarity or abnormality of a discrepancy between a predicted and obtained score are subject to sampling error. They will therefore have an intuitive appreciation that, for example, less confidence should be placed in them when a regression equation was built using a small sample. However, the advantage of the procedures to be outlined is that they quantify the degree of confidence that should be placed in these estimates of abnormality. Furthermore, as will be shown, other factors influence the precision of these estimates so that relying on an informal internal model as a basis for judging how much confidence to place in them is not feasible.

The methods required, particularly those for obtaining confidence limits and those used when there are multiple predictor variables, are complex but they have all been implemented in accompanying computer programs. As a result, clinicians or researchers need never carry out the computations involved. A further positive feature of these methods is that they can be used with existing (i.e., previously published) regression equations, provided that basic statistics for the equations are available.

Using the Standard Error of Estimate to Evaluate Discrepancies

Before setting out the methods, consideration will be given to a commonly used alternative means of drawing inferences concerning discrepancies between predicted and obtained scores (for examples, see Crawford, Moore, & Cameron, 1992; Knight & Shelton, 1983; McSweeny, Naugle, Chelune, & Lüders, 1993; Paolo, Ryan, Troster, & Hilmer, 1996; Temkin et al., 1999.) This method makes use of what, in the psychological literature, is referred to as the *standard error of estimate* ($s_{Y \cdot X}$); in mainstream statistics, the term *residual standard deviation* is more commonly employed. 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. As such, it reflects the precision of our estimation procedure. The equation for the standard error of estimate for bivariate cases (i.e., when there is only a single predictor variable) is

$$s_{Y \cdot X} = s_Y \sqrt{(1 - r^2)\frac{N-1}{N-2}}, \tag{1}$$

where s_Y is the standard deviation of the criterion variable, r^2 is the squared correlation between the predictor and criterion variables, and N is the size of the sample used to build the equation.

The method involves dividing the discrepancy between an individual's obtained and predicted scores by $s_{Y \cdot X}$, treating the result as a standard normal deviate (i.e., z), and referring this z to a table of the areas under the normal curve. For example, suppose that a regression equation predicting retest scores on a memory test from scores on initial testing had a $s_{Y \cdot X}$ of 5, that a patient's actual score at retest (Y) was 80, and the score predicted from her or his initial score ($\hat{Y}$) was 89. Further, suppose that a researcher or clinician wished to test the directional hypothesis that the patient's obtained score was lower than that predicted by the equation (i.e., the researcher hypothesized that performance will have deteriorated and hence employed a one-tailed test). Dividing the discrepancy of -9 by 5 gives a result of -1.8, and referring to a table of the normal curve reveals that the p value associated with this z is 0.036. Therefore, it would be concluded that the patient's performance at retest was significantly below expectations and that there had been a genuine deterioration in memory functioning.

There are a number of factors that are ignored when this approach is employed. First, as noted above, $s_{Y \cdot X}$ allows us to make statements about the sample used to build the equation, but in this situation it is being used to draw inferences concerning an individual who is not a member of the sample. Second, by using a standard normal deviate for inferential purposes, the sample used to build the equation is treated as though it were the population of interest; that is, the statistics obtained from the sample are treated as population parameters rather than sample statistics (Crawford &

Howell, 1998a). Thus, technically, results should be referred to a *t* distribution on $N - 2$ *df* rather than the standard normal distribution (two degrees of freedom are lost in calculating the slope and intercept for the regression equation). Furthermore, the method does not deal with the fact that the error associated with predicting *Y* from *X* will be greater at more extreme values of *X* than at values close to the mean of *X*. This occurs because of error in estimating the population regression slope from the sample regression slope. To visualize this effect, imagine rotating the regression line around the mean of *X* (Crawford & Howell, 1998b). The upshot is that the method described should be associated with inflated Type I errors. That is, the number of cases incorrectly classified as exhibiting a significant discrepancy will be higher than the specified error rate. The method to be outlined deals with all of these shortcomings.

Study 1

In Study 1, the method for the significance test and corresponding point estimate of the abnormality of a patient's discrepancy are derived for the bivariate case (i.e., for regression equations in which there is only one predictor variable) and the vector case (i.e., where there are multiple predictor variables). Monte Carlo simulations are performed to assess whether the method achieves control of the Type I error rate at various values of *N* and of ρ_{XY} (the population correlation between *X* and *Y*). In the present context, a Type I error would occur if a member of the population was misclassified as not having been drawn from the population. For comparative purposes, we also examine Type I errors for the commonly employed method that uses $s_{Y \cdot X}$ for inferential purposes.

The statistically sophisticated reader may consider that running this first set of simulations is unnecessary because theory would predict that the use of the $s_{Y \cdot X}$ method will fail to control Type I errors whereas the method to be outlined should achieve adequate control. However, we had two reasons for conducting them. First, because the use of the $s_{Y \cdot X}$ method is very widespread, it would be useful to quantify the inflation of the Type I error rate across a range of scenarios so that neuropsychologists can make informed decisions on its use. Second, for those readers who have limited interest in statistical theory but nevertheless need to use quantitative methods in their research or practice, an empirical demonstration of the control achieved by the two methods may be more convincing than appeal to theory alone.

Method

Derivation of the method. The regression equation relating *X* and *Y* is

$$Y = a + bX + \varepsilon,$$

where ε is the random error. We assume errors (ε) are normally distributed with a variance (σ^2) that does not depend on *X*, so

$$Y \sim N(a + bX, \sigma^2).$$

When there is only a single predictor variable, the standard error of a predicted score for a new case (Crawford & Howell, 1998b; Howell, 2002) is

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

where $\bar{X}$ and s_X^2 are the sample mean and variance of *X* for the controls, respectively; X_0 is the patient's score on the predictor; and the remaining terms have been defined previously. It can be seen that the equation for this standard error incorporates $s_{Y \cdot X}$ but that it will always be larger than the latter quantity. It can also be seen that as *N* increases the magnitude of s_{N+1} will decrease. In addition, it can be seen that the s_{N+1} will increase in magnitude the further the patient's score on *X* is from the mean of *X*. Put

$$\frac{Y_0^* - \hat{Y}}{s_{N+1}}\,, \qquad (3)$$

where Y_0^* is the obtained score for the patient and $\hat{Y}$ is the predicted score, to yield a standardized discrepancy between the predicted and obtained score. Under the null hypothesis, this quantity will have a *t* distribution on $N - 2$ *df*. Thus, for a specified level of alpha (e.g., 0.05), one can test whether there is a statistically significant difference (one-tailed) between the predicted score and the obtained score. In addition, multiplying the resultant *p* value by 100 will provide a point estimate of the percentage of the population that would obtain a discrepancy more extreme than that observed for the patient. Note that, if a two-tailed test is required, alpha would be divided by 2. For example, the appropriate critical value would be that corresponding to the 2.5th rather than 5th percentile point of the relevant *t* distribution.

Turning now to the vector case, the regression equation relating $\underline{X} = (X_1, X_2, \ldots, X_k)'$ and *Y* is

$$Y = a + b_1X_1 + b_2X_2 + \ldots b_kX_k + \varepsilon = a + \underline{b}'\underline{X} + \varepsilon, \qquad (4)$$

where ε (as before) is the random error and $\underline{b}$ and $\underline{X}$ are $k \times 1$ vectors. Assuming normality,

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

We suppose a sample of size *N* gives least-squares estimates $\hat{a}$ and $\underline{\hat{b}}$ of *a* and $\underline{b}$, and $s^2_{Y \cdot \underline{X}}$ as the estimate σ^2. Then,

$$(N - k - 1)s^2_{Y \cdot \underline{X}}/\sigma^2 \sim \chi^2_{N-k-1}. \qquad (5)$$

The computations required for the standard error of a predicted score for a new individual in the vector case are more complex than in the bivariate case. This standard error can be expressed in a number of different but equivalent forms; here, we adapt the form used by Cohen et al. (2003, equation A1.8):

$$s_{N+1} = s_{Y \cdot \underline{X}} \sqrt{1 + \frac{1}{N} + \frac{1}{N-1}\Sigma r^{ii} z_{io}^2 + \frac{2}{N-1}\Sigma r^{ij} z_{io} z_{jo}}, \qquad (6)$$

where r^{ij} identifies off-diagonal elements of the inverted correlation matrix for the *k* predictor variables, r^{ij} identifies elements in the main diagonal, and $z_0 = (z_{10}, \ldots, z_{k0})'$ identifies the patient's scores on the predictor variables in *z* score form. Note that Cohen et al. (2003) define $z_{i0} = n(x_{i0} - \bar{x}_i)/\Sigma(x_{ij} - \bar{x}_i)^2$ while we use the more usual form[1] $z_{i0} = (n - 1)(x_{i0} - \bar{x}_i)/\Sigma(x_{ij} - \bar{x}_i)^2$. The first summation in Equation 6 is over the

[1] This difference stems from the fact that, in earlier editions of Cohen's text, the equations for regression analysis used *n* in the denominator for standard deviations rather than *n* – 1. Although Cohen et al. (2003) state in the third edition (p. 17) that they have moved to using *n* – 1 (which is the standard convention and is used in our equations), there are still some hangovers from earlier editions, of which their equation A1.8 is one example.

k diagonal elements and the second is over the $k(k - 1)/2$ off-diagonal elements below (or above) the diagonal.

When the standard error for the vector case is substituted in the denominator of Equation 3, then, under the null hypothesis, the resultant quantity will have a t distribution on $N - k - 1$ *df*. Thus, as in the bivariate case, one can test for a significant difference between the obtained and predicted score and achieve an estimate of the abnormality of the discrepancy.

Evaluation of the method by Monte Carlo simulation. Monte Carlo simulations were run on a PC and implemented in Borland Delphi (Version 4). In the interest of brevity, and because this first set of simulations was designed primarily for didactic purposes, we report results for the bivariate case only. Simulations for the vector case yielded the same pattern of results and are available from the first author on request.

The first simulation examined the overall Type I error rate for the present method and the $s_{Y \cdot X}$ method. The algorithm ran3.pas (Press, Flannery, Teukolsky, & Vetterling, 1989) was used to generate uniform random numbers (between 0 and 1), and these were transformed by the polar variant of the Box–Muller method (Box & Muller, 1958). The Box–Muller transformation generates pairs of normally distributed observations and, by further transforming the second of these, it is possible to generate observations from a bivariate standard normal distribution with a specified correlation (e.g., see Kennedy & Gentle, 1980).

The simulation was run with six different values of N (the size of the sample used to build the regression equation): 5, 10, 20, 50, 100, and 200. For each of these values of N, 1,000,000 samples of $N + 1$ were drawn from one of five bivariate normal distributions in which ρ_{XY} was set at either 0.4, 0.6, 0.7, 0.8, or 0.9 (without loss of generality, the population means and variances were set equal to 0 and 1, respectively). Thus, a total of 30 million individual Monte Carlo trials were run.

In each trial, the first N pairs of observations were taken as the X and Y scores of the sample and were used to build the regression equation. The $(N + 1)$th pair was taken as the scores of the individual control case. Note that the concern is with Type I errors in this scenario; therefore, the individual case is not a patient but a member of the same population as the sample used to build the equation and the interest is in the percentage of such cases that are misclassified. The significance test developed in the present paper was then applied to the scores of the control case. That is, the discrepancy between the predicted and obtained score was divided by s_{N+1} and the result evaluated against a t distribution on $N - 2$ *df*. A one-tailed test was employed using a specified Type I error rate of 5%. A one-tailed test was chosen because neuropsychologists will normally have a directional hypothesis concerning discrepancies between predicted and obtained scores. For comparative purposes, the discrepancy was also divided by $s_{Y \cdot X}$ and the result (z) recorded as significant if it exceeded -1.645 (i.e., a one-tailed test with a specified error rate of 5% was also applied to this latter statistic).

The second simulation examined Type I errors for the two methods as a function of the extremity of the score on the predictor (X) variable. The simulation method was identical in most respects to the first simulation (i.e., a series of samples of size N were drawn from one of five bivariate normal distributions and used to build regression equations). Type I errors were examined for the same six values of N examined in the first simulation. However, unlike the first simulation, in which the $(N + 1)$th case was drawn randomly from the same bivariate normal distribution as controls, the X value for the case (X_0) was fixed at one of five values: 0, -0.5, -1.0, 1.5, and 2.0 (because the distribution of the parent population for X is standard normal, these values are *SD* units). The Y value was generated using the equation

$$Y = \rho_{XY}X + \sqrt{1 - \rho_{XY}^2}z,$$

where X_0 is one of the fixed values referred to above, $\sqrt{1 - \rho_{XY}^2}$ is the population residual standard deviation, and z is a random draw from a standard normal distribution. A million trials were run for each combination of N, ρ_{XY}, and X_0; thus, in the second simulation, 150 million Monte Carlo trials were run in total. To reiterate, unlike the first simulation, which was designed to examine the overall Type I error rate, this second simulation was designed to quantify the Type I error rate as a function of the extremity of the scores on the independent (X) variable.

Results and Discussion

The basic pattern of results from the first simulation can be clearly appreciated by examining Figure 1. Because the simulations varied both the size of the control sample (N) and the population correlation between tasks (ρ_{XY}), representing all these results in a single figure would be messy. We have therefore presented the results for $\rho_{XY} = 0.7$ only; the same pattern would be obtained for other values of ρ_{XY}. The full results from the first simulation are presented in Table 1.

It can be seen from Table 1 that Type I errors are not controlled when the commonly employed method (based on $s_{Y \cdot X}$) is used to test for a significant discrepancy between an individual's predicted and obtained score. The error rates range from a high of 14.24% for a ρ_{XY} of 0.9 and an N of 5 to a low of 5.16% for a ρ_{XY} of 0.9 and an N of 200 (the small differences in error rates across levels of ρ_{XY} do not exceed those expected from Monte Carlo variation). It can be seen that the error rates are very inflated with small to moderate Ns, although with Ns of 100 and above the degree of inflation is modest.

In contrast to the results for the $s_{Y \cdot X}$ method, use of the method derived in the present paper achieves excellent control over Type I errors across all values of N and ρ_{XY}; that is, the error rates cleave closely to the specified rate of 5%. The error rates range from a

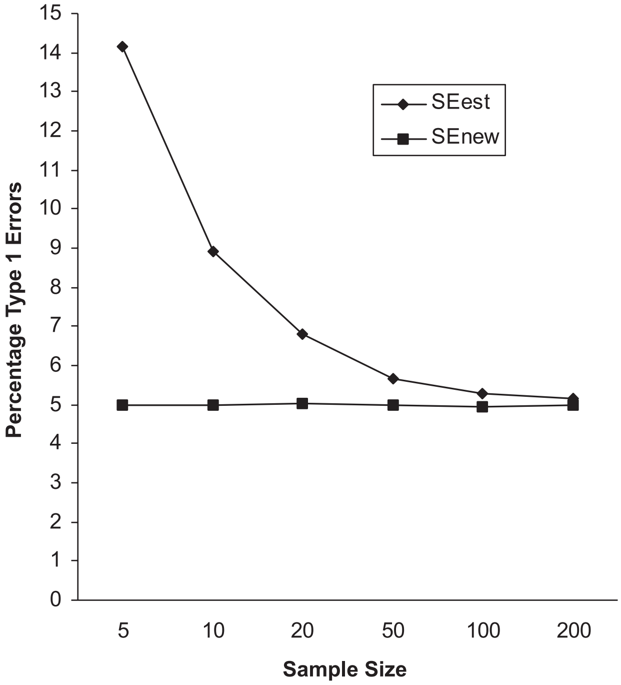

Figure 1. Monte Carlo simulation results: percentage of Type I errors for the two methods of testing for a significant discrepancy between predicted and obtained scores (in this example $\rho_{XY} = 0.7$). SEest = the standard error of estimate; SEnew = the standard error for a new individual case.

Table 1
Monte Carlo Simulation Results: Percentage of Type I Errors for the Two Methods of Testing for a Significant Discrepancy Between Predicted and Obtained Scores

	$s_{Y \cdot X}$ method					s_{N+1} method				
N	0.4	0.6	0.7	0.8	0.9	0.4	0.6	0.7	0.8	0.9
5	14.24	14.16	14.14	14.18	14.24	4.99	5.00	4.97	5.01	5.02
10	8.92	8.89	8.91	8.90	8.95	5.03	5.01	5.00	4.98	5.02
20	6.79	6.83	6.82	6.81	6.85	4.98	5.01	5.03	5.00	5.05
50	5.72	5.68	5.67	5.67	5.69	5.03	5.01	4.98	4.99	5.01
100	5.32	5.37	5.30	5.29	5.29	4.98	5.03	4.96	4.97	4.97
200	5.18	5.21	5.17	5.18	5.16	5.01	5.04	5.00	5.01	5.00

high of 5.03% to a low of 4.97% (the small deviations from the specified error rate of 5% are within the range expected from Monte Carlo variation).

Results from the second simulation, in which the error rates for specific (i.e., fixed) values of X_0 were examined, are presented in Table 2. It can be seen that the error rates for the $s_{Y \cdot X}$ method become more inflated with increasingly extreme values of X_0. For example, for an N of 20 and ρ_{XY} of 0.7, the error rates range from 6.32% when the score on the predictor variable is at the mean to 8.22% when X_0 is 2 *SD*s from the mean. In contrast, it can be seen that when the method described in the present paper is applied the error rate is under control regardless of the values of X_0.

Table 2
Percentage of Type I Errors for the Two Methods of Testing for a Significant Discrepancy Between Predicted and Obtained Scores as a Function of the Extremity of the Predictor Variable Score

	$s_{Y \cdot X}$ method			s_{N+1} method		
N	0.4	0.7	0.9	0.4	0.7	0.9
$X_0 = 0$						
5	12.16	12.17	12.16	5.01	5.01	5.00
10	7.87	7.87	7.88	4.99	5.00	4.99
20	6.32	6.32	6.30	5.01	5.00	5.01
50	5.50	5.50	5.50	4.99	5.01	5.00
100	5.23	5.25	5.23	4.98	5.01	5.00
200	5.11	5.13	5.12	4.98	5.01	5.00
$X_0 = 0.5$						
5	12.77	12.80	12.86	4.98	5.00	4.99
10	8.19	8.13	8.14	5.02	4.97	4.99
20	6.40	6.45	6.42	4.97	5.01	5.02
50	5.57	5.55	5.55	5.01	4.99	4.99
100	5.28	5.26	5.23	5.00	4.99	5.01
200	5.10	5.12	5.14	4.96	4.99	5.00
$X_0 = 1.0$						
5	14.61	14.56	14.65	5.02	4.99	4.97
10	8.96	9.04	9.04	4.99	5.04	4.97
20	6.82	6.84	6.82	5.02	5.03	5.00
50	5.69	5.67	5.69	5.01	5.01	4.99
100	5.31	5.32	5.36	4.99	5.01	5.00
200	5.15	5.15	5.17	4.99	4.98	5.01
$X_0 = 2.0$						
5	19.37	19.41	19.40	4.97	4.98	5.00
10	11.94	11.84	11.77	5.06	5.01	4.97
20	8.19	8.23	8.17	5.00	5.03	4.97
50	6.21	6.23	6.19	4.99	5.01	4.98
100	5.57	5.61	5.56	4.98	5.01	4.97
200	5.31	5.31	5.29	5.02	5.01	4.99

It was noted that the significance tests on regression discrepancies also simultaneously provide point estimates of the abnormality or rarity of an individual's discrepancy. It follows from the results of the present simulations that when the $s_{Y \cdot X}$ method is used the point estimates will exhibit a systematic bias; they will exaggerate the abnormality or rarity of the individual's score. It also follows from the simulations that the method presented here will not suffer from this systematic bias.

As a specific example, suppose that a clinician or researcher used a regression equation built with a healthy control sample of $N = 20$, that the ρ_{XY} was 0.7, and that a patient's score on X was 2 *SD*s below the mean. Suppose also that the $s_{Y \cdot X}$ method yielded a z of -1.645. It would be concluded that only 5% of the control population would obtain a larger discrepancy than that observed for the patient. However, as noted, it can be seen from Table 2 that it is to be expected that 8.22% of the population with this value of X would exceed the individual's discrepancy; that is, the rarity of the discrepancy has been exaggerated by the $s_{Y \cdot X}$ method.

Study 2

In Study 2, the method for obtaining confidence limits on the abnormality of a patient's discrepancy is derived. That is, in Study 1 the method for obtaining a point estimate of the percentage of the population that would obtain a more extreme discrepancy was derived and evaluated. The present study aims to derive and evaluate a method for obtaining 95% confidence limits on this percentage. The method used is an extension to those developed by Crawford and Garthwaite (2002) and is based on noncentral t distributions. An excellent overview of the use of noncentral t distributions in obtaining confidence limits for other quantities is provided by Cumming and Finch (2001).

Method

Derivation of the method for confidence limits on the abnormality of discrepancies. The confidence intervals are derived from a noncentral 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 noncentrality parameter and effects the shape and skewness of the distribution.

For the bivariate case, as in Study 1, for a patient we observe values X_0 and Y_0^*. When the patient's predicted score ($\hat{Y}_0$) exceeds the obtained score (Y_0^*), let $P^* = \Pr(Y_0 < Y_0^*) \cdot 100$, where Y_0 is the value of a control whose X value is X_0. Then P^* is the percentage of people with the same X value as the patient who would have a Y value below Y_0^*. When the patient's obtained score (Y_0^*) exceeds the predicted score ($\hat{Y}_0$), then let $\text{P}^* = \Pr(Y_0 > Y_0^*) \cdot 100$, where, as before, Y_0 is the value of a control whose X value is X_0. Then P^* is the percentage of people with the same X value as the patient who would have a Y value above Y_0^*. To show which of these two situations holds, we let τ be an indicator where $\tau = 1$ if $\hat{Y}_0$ exceeds Y_0^*, and $\tau = -1$ if $\hat{Y}_0$ is less than or equal to Y_0^*.

We require a $100(1 - \alpha)\%$ confidence interval for P^* when a sample of size N gives estimates $\hat{a}$, $\hat{b}$, and $s^2_{Y\cdot X}$ for a, b, and σ^2. Let $\hat{Y}_0$ be the predicted value of a patient whose X value is X_0, where the prediction is based on the sample used to build the equation. The standard error of $\hat{Y}_0$ is

$$\sigma\sqrt{\frac{1}{N} + \frac{(X_0 - \bar{X})^2}{(N-1)s_X^2}}. \tag{7}$$

This equation represents the standard deviation of the errors in predicting average scores on Y for the population from a sample. Now define

$$\theta = \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{(N-1)s_X^2}. \tag{8}$$

Then, we have

$$\hat{Y}_0 \sim N(a + bX_0, \sigma^2\theta).$$

Put

$$c = \frac{Y_0^* - \hat{Y}_0}{s_{Y\cdot X}}, \tag{9}$$

where $s^2_{Y\cdot X}$ is the estimate of σ^2.

As an aside, it was noted that Equation 3 provides a standardized difference between the obtained and predicted score (the raw difference is divided by s_{N+1}). It can be seen that in Equation 9 the raw difference between obtained and predicted scores is divided by the standard error of estimate ($s_{Y\cdot X}$): c therefore also represents a standardized difference. However, in this latter case, the imprecision in estimating the regression line from sample data is ignored. This feature makes it unsuitable as the basis for a significance test on the discrepancy (although, as noted earlier, it is widely used for this purpose in neuropsychology) or as a means of obtaining a point estimate of the abnormality of the discrepancy. However, this same feature makes it suitable as an index of effect size so that it has value beyond its current role as an intermediate step in finding the required confidence limits.

Returning once more to the derivation for finding confidence limits on the abnormality of the difference, let

$$c^* = \frac{Y_0^* - a - bX_0}{\sigma}. \tag{10}$$

Then c is an estimate of c^*. Percentage points of a normal distribution determine P^* from c^*, so a $100(1 - \alpha)\%$ confidence interval for c^* will yield the required confidence interval for P^*. Now,

$$\frac{c}{\sqrt{\theta}} = \frac{(a + bX_0 - \hat{Y}_0)/\sqrt{\sigma^2\theta} + (Y_0^* - a - bX_0)/\sqrt{\sigma^2\theta}}{\sqrt{s^2_{Y\cdot X}/\sigma^2}}, \tag{11}$$

and

$$(N-2)s^2_{Y\cdot X}/\sigma^2 = \chi^2_{N-2}.$$

Hence, $c/\sqrt{\theta}$ has a noncentral t distribution with noncentrality parameter $\delta = c^*/\sqrt{\theta}$ and $N - 2$ df. The $100(\alpha / 2)\%$ and $100(1 - \alpha / 2)\%$ points of this distribution will depend on the value of δ. Let δ_L denote the value of δ for which the $100(1 - \alpha / 2)\%$ point is $\tau c/\sqrt{\theta}$. Note that $\tau c/\sqrt{\theta}$ is negative regardless of whether $\hat{Y}_0$ is greater or less than Y_0^* because of the indicator τ. Similarly, let δ_U denote the value of δ for which the $100(\alpha / 2)\%$ point is $\tau c/\sqrt{\theta}$. Then $(\delta_L\sqrt{\theta}, \delta_U\sqrt{\theta})$ is a $100(1 - \alpha)\%$ confidence interval for τc^*. Define h_1 and h_2 by

$$h_1 = \Pr(Z < \delta_L\sqrt{\theta}) \times 100, \tag{12}$$

and

$$h_2 = \Pr(Z < \delta_U\sqrt{\theta}) \times 100, \tag{13}$$

where Z is the standard normal variate [that is, $Z \sim N(0, 1)$]. Then a $100(1 - \alpha)\%$ confidence interval for P^* is (h_1, h_2).

For some purposes, it would be useful to have a one-sided confidence limit. A one-sided limit can be obtained by modifying the above procedure such that, if the lower confidence limit is required, δ_L denotes the value of δ for which the $100(1 - \alpha)\%$ point is $c/\sqrt{\theta}$. Similarly, if the upper confidence limit is required, δ_U would denote the value of δ for which the $100(\alpha)\%$ point is $c/\sqrt{\theta}$.

Turning now to the vector case, for a patient we observe values $\underline{X}_0$ and Y_0^*. Put

$$\hat{Y}_0 = \hat{a} + \underline{\hat{b}}'\underline{X}_0, \tag{14}$$

where $\hat{a}$ and $\hat{b}$ are the estimates of the regression coefficients in Equation 4. As before, if $\hat{Y}_0$ exceeds Y_0^*, then let $P^* = \Pr(Y_0 < Y_0^*) \cdot 100$, where Y_0 is the value of a control whose X values are now the vector $\underline{X}_0$. If $\hat{Y}_0$ is less than or equal to Y_0^*, let $P^* = \Pr(Y_0 > Y_0^*) \cdot 100$. Also, put

$$c = \frac{Y_0^* - \hat{Y}_0}{s_{Y\cdot \underline{X}}}. \tag{15}$$

(Note that, as in the bivariate case, c is an index of effect size). Now put

$$\theta = \frac{1}{N} + \frac{\Sigma r^{ii} z_{io}^2 + 2\Sigma r^{ij} z_{io} z_{jo}}{N-1}, \tag{16}$$

where all terms are as defined in Study 1. In the Appendix, which is available on the Web at http://dx.doi.org/10.1037/0894-4105.20.3.259.supp, it is shown that

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

Following similar lines to those for one explanatory variable, let δ_L denote the value of δ for which $\tau c/\sqrt{\theta}$ is the $100(1 - \alpha / 2)\%$ point of a noncentral t distribution on $(N - k - 1)$ df, where τ is the indicator function defined earlier. Similarly, let δ_U denote the corresponding value for the $100(\alpha / 2)\%$ point. Define h_1 and h_2 as in Equations 12 and 13. Then a $100(1 - \alpha)\%$ confidence interval for P^* is (h_1, h_2).

Evaluation of the method of setting confidence limits by Monte Carlo simulation. To evaluate the confidence limits, a population was specified as was the precise level of abnormality for a given pair of X_0 and Y_0^*. That is, values of c^* in Equation 10 were specified, thereby obtaining values for P^*. Then, on each Monte Carlo trial, a sample was drawn from this population, the sample statistics required to build a regression equation and to obtain the confidence limits were calculated, and the program recorded whether the limits include the specified (i.e., true) level of abnormality

(P*). If the method is sound, the confidence limits should include P* on 95% of Monte Carlo trials.

As in Study 1, we present results only for the bivariate case; results for the vector case can be obtained from the first author. Three different values of N were used: 5, 20, and 200. In addition, we employed three values of X_0 (0, 1, and 2 *SD*s below the population mean) and three values for the abnormality (d) of the discrepancy between obtained and predicted scores: 1%, 5%, and 25%. For each of these combinations of N, X_0, and d, 100,000 samples of size N were drawn from one of three bivariate normal distributions in which the ρ_{XY} was set at either 0.4, 0.7, or 0.9. Thus, a total of 8.1 million individual Monte Carlo trials were run. A smaller number of values for N, ρ_{XY}, and X_0 and trials per condition were employed than in Study 1 because the present simulation was much more computationally intensive.

For computational convenience and without loss of generality, the population means and *SD*s of X and Y were set at 0 and 1, respectively. The values for Y_0^* were generated using the equation

$$Y_0^* = \rho_{XY}X_0 + \sqrt{1 - \rho_{XY}^2}z_a, \tag{17}$$

where, as noted, X_0 took one of three values (0, 1, or 2); z_a, a standard normal deviate, took one of three values corresponding to the required level of abnormality (2.3263, 1.6449, or 0.6745); and $\sqrt{1-\rho_{XY}^2}$ is the population residual standard deviation. Note that the intercept does not appear in Equation 17 because it was zero.

Results and Discussion

The results of the Monte Carlo simulation evaluating the confidence limits are presented in Table 3. If the method for finding these limits is valid, then, as noted, they should capture the true level of abnormality 95% of the time, regardless of the sample size (N), the correlation between X and Y (ρ_{XY}), the level of extremity of the score on the predictor variable (X_0), and the specified level of abnormality of the discrepancy. It can be seen from Table 3 that the limits achieve this (the small deviations from 95% are within the bounds expected from Monte Carlo variation). Furthermore, in each condition, the percentage of trials in which the true abnormality (P*) lay below the lower limit was closely balanced by the percentage in which P* lay above the upper limit (in the interest of brevity, these figures are not reported in the table but are available on request).

As noted, the confidence limits derived in the present study quantify the degree of uncertainty surrounding the estimate of the abnormality (i.e., rarity) of a discrepancy between an obtained and predicted score. Take, for example, the case of limits that range from 0.002% to 1.3%. The user can be 95% confident that the percentage of the population that would exceed the discrepancy observed for the patient lies within these limits; that is, in this example, the user can be confident that the patient's discrepancy is unusual.

The width of these limits is determined by the size of the sample used to generate the equation, by the extent to which an individual's X score (X_0) deviates from the mean of X for the sample, and by the magnitude of the correlation between X and Y. Table 4 serves to illustrate the former of these two effects and also illustrates the relationship between the limits and the point estimates (the latter were obtained using the method derived in Study 1). The confidence limits and point estimates in Table 4 were generated by setting the means and standard deviations of X and Y at 0 and 1, respectively; the correlation between X and Y at 0.7; and the raw difference between the predicted and obtained scores at -1.4. Three values of X_0 (0, -1, and -2; these values are standard deviation units) and four values of N (5, 20, 200, and 1000) were employed.

It can be seen from Table 4 that the limits are very wide when small samples are used to build regression equations. For example, for an N of 20 and an X_0 of one standard deviation below the mean, the limits range from 0.28% to 15.55%. However, even with larger Ns the limits are not insubstantial. For example, with an X_0 of one standard deviation below the mean, the limits range from 1.29% to 4.65% for an N of 200, and from 1.87% to 3.32% for an N of 1000. Thus, these limits provide a useful means of quantifying the uncertainty associated with estimates of abnormality across a range of situations in which regression is employed to draw inferences concerning an individual patient. That is, it applies in contexts from neuropsychological single-case studies, in which a modestly sized sample of other patients or healthy control participants has been recruited for comparative purposes, to clinical situations, in which inferences are made using a regression equation built from a large standardization sample. It can also be seen

Table 3
Monte Carlo Simulation Evaluating the Method of Setting Confidence Limits on the Abnormality of Discrepancies Between Predicted and Obtained Scores

	$\rho_{XY} = 0.4$			$\rho_{XY} = 0.7$			$\rho_{XY} = 0.9$		
N	1%	5%	25%	1%	5%	25%	1%	5%	25%
$X = 0$									
5	94.95	94.99	94.88	95.12	95.08	94.94	95.00	95.02	94.90
20	95.02	95.00	95.02	95.05	94.96	94.94	95.05	94.98	95.17
200	95.05	94.98	95.01	94.94	95.06	95.04	94.96	95.01	94.97
$X = 1$									
5	94.93	95.13	95.09	94.94	95.01	95.09	95.10	94.92	95.01
20	95.09	94.88	95.02	94.96	94.99	95.04	95.08	95.06	94.98
200	94.97	94.92	94.96	95.10	95.00	94.92	95.02	94.96	95.03
$X = 2$									
5	95.12	95.03	94.93	94.93	94.99	95.00	95.11	94.99	95.07
20	95.10	94.94	94.99	94.97	95.04	94.89	94.96	94.97	94.91
200	94.89	95.10	95.05	95.00	95.07	95.01	94.95	95.03	94.98

Table 4
Examples Illustrating Factors Effecting the Width and Degree of Asymmetry of 95% Confidence Limits on the Abnormality of Discrepancies Between Predicted and Obtained Scores

	$X = 0$				$X = -1$				$X = -2$			
N	Lower	Point	Upper	(MidCL)	Lower	Point	Upper	(MidCL)	Lower	Point	Upper	(MidCL)
5	0.06	10.95	45.37	(22.71)	0.02	12.67	57.99	(29.00)	0.02	16.77	81.63	(40.82)
20	0.39	3.95	12.69	(6.54)	0.28	4.29	15.55	(7.91)	0.11	5.32	23.80	(11.96)
200	1.42	2.63	4.30	(2.86)	1.29	2.66	4.65	(2.97)	1.02	2.74	5.61	(3.32)
1000	1.95	2.52	3.19	(2.57)	1.87	2.53	3.32	(2.59)	1.69	2.55	3.63	(2.66)

Note. The entries are percentages. These examples are all based on means and *SD*s of 0 and 1, respectively; for *X* and *Y*, an r_{XY} of 0.7; and a raw difference between predicted and obtained scores of −1.4. MidCL records the midpoint between the upper and lower limits and is included only to assist in demonstrating that the confidence limits are not symmetrical around the point estimate.

from Table 4 that the limits widen as the score on the predictor becomes more extreme; this is particularly marked with small to moderate *N*s.

The limits in Table 4 are two-sided limits, but one-sided limits may often be more in keeping with a neuropsychologist's assessment aims. For example, if a researcher or clinician is concerned about how common an individual's discrepancy might be, but uninterested in whether it may be even more unusual than the point estimate indicates, then a one-sided upper limit is more appropriate (the program described below calculates one- or two-sided limits). For example, rather than state, "There is 95% confidence that the proportion of people who have a discrepancy as large as the patient's is between 0.28% and 15.55%," we might prefer to state, "There is 95% confidence that the proportion of people who have a discrepancy as large as the patient's is less than 12.51%."

Table 4 illustrates another feature of the confidence limits: They are nonsymmetrical around the point estimate; the lower limit (which must exceed 0) is nearer to the point estimate than is the upper limit. This asymmetry occurs because the limits follow a noncentral *t* distribution. To help the reader appreciate this characteristic, Table 4 records the midpoint between the upper and lower limits (these figures are bracketed in the table to make it clear that they are for illustrative purposes only; i.e., it is not suggested that this quantity should be employed when using the limits).

Figure 2 is designed to provide further insight into the nature of the noncentral *t* distributions used to obtain the confidence limits. As noted, to obtain the lower and upper limits on the abnormality of the discrepancy requires the use of two noncentral *t* distributions. Figure 2 graphs these pairs of distributions when the sample size is either small ($n = 10$; $df = 8$) or modest ($n = 30$; $df = 28$) in size, and the standardized discrepancy *c* (Equation 9) between predicted and obtained scores, which determines the noncentrality parameters, is either modest (−0.5) or large (−2.0) in magnitude. It can be seen that the distributions are more asymmetric when *N* is small and the noncentrality parameter is large (i.e., when the discrepancy is extreme).

Given the complex determinants of the confidence limits, it will be appreciated that it is not feasible for neuropsychologists to rely on an informal model when attempting to determine the level of confidence that should be placed in an estimate of the abnormality of a discrepancy between predicted and obtained scores.

General Discussion

Utility of the Methods

Because the $s_{Y \cdot X}$ method performs reasonably well when the *N* used to build an equation is large, it could be argued that, in such circumstances, it should be used in preference to the present methods because it is simpler. However, it remains the case that the former method is technically incorrect; that is, the Type I error rate will exceed the specified rate and the estimate of abnormality will be exaggerated, even if these effects are relatively modest with moderate to large *N*s. Second, if the programs accompanying this paper are employed (see later section), the present methods are, if anything, faster and easier to use in practice. Furthermore, the $s_{Y \cdot X}$ method cannot provide confidence limits on the abnormality of a patient's discrepancy and, as was shown, there is still appreciable uncertainty attached to the estimates of abnormality even when *N* is large.

Therefore, the present methods have the advantage over the $s_{Y \cdot X}$ method in that they remain valid regardless of the sample size used to build the equation and the extremity of the patient's score on the predictor variable. Furthermore, the provision of confidence limits is (a) in keeping with APA guidelines on statistics and test standards, (b) serves the general purpose of reminding neuropsychologists that the results they obtain when using regression equations are fallible, and (c) serves the specific purpose of quantifying this fallibility.

Significance Test and Confidence Limits Contrasted

Confidence limits can often be used as an alternative means of testing for statistical significance. For example, if an independent samples *t* test is used to test the null hypothesis that the difference between two group means is zero, we can reject this hypothesis if the resultant probability is less than our specified level of alpha (say, the 5% level). The same conclusion can be reached if the 95% confidence limits on the difference between means does not include zero.

In the present case, however, it is very important to appreciate that the significance test and confidence limits address different aspects of a patient's data. For example, let us suppose that a clinician or researcher wants to test the directional hypothesis that a patient's obtained score is significantly lower than the patient's predicted score. The significance test is applied and it is

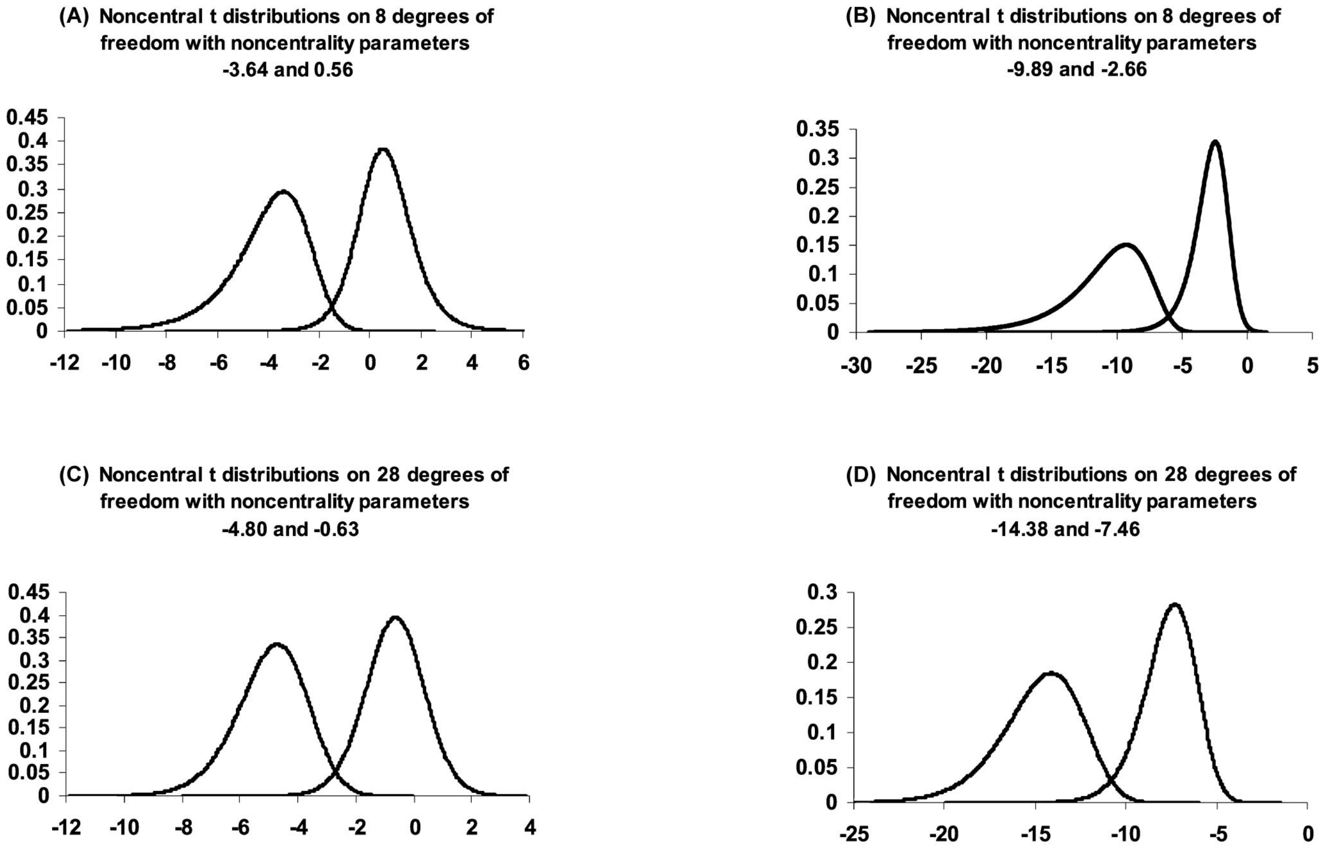

Figure 2. Plots of the pairs of noncentral *t* distributions used to form the confidence limits on the abnormality of the discrepancies between predicted and obtained scores. The plots illustrate the effect on the shape of these distributions of varying the sample size (small = A and B, or moderate = C and D) and the magnitude of the discrepancy (moderate = A and C, or large = B and D).

found that the null hypothesis can be rejected; that is, the one-tailed *p* value was < .05. As noted, the null hypothesis is that the patient's discrepancy was an observation from the control population.

Suppose also that, as will commonly be the case particularly when *N* is modest, the one-sided 95% upper limit on the abnormality of the discrepancy is considerably larger than 5% (say, 11%). This latter result in no way invalidates the result of the significance test; 100*p*% is the point estimate of the percentage of individuals who would have a discrepancy as low as the patient's score, and this estimate will be comfortably within the confidence interval for that percentage. However, the upper confidence limit will be greater than 100*p*% and the difference will be substantial if *p* is known only very imprecisely, as will usually be the case when *N* is small. As noted, the confidence limits serve the useful purpose of quantifying this imprecision and are also a general reminder that the results of any regression analysis in the individual case are fallible (i.e., they help us avoid reifying our results). Note, however, that when a significant result is obtained and the upper confidence limit is of small magnitude, then the researcher or clinician can be confident that the individual's discrepancy is very unusual.

Present Method Contrasted With Reliable Change Indices

It has been stressed that the present method can assist in the process of identifying and quantifying change in neuropsychological functioning in the individual case. Because reliable change indices (e.g., Jacobson & Truax, 1991) are used for a similar purpose, it is appropriate to compare and contrast these two approaches. It will not be possible to conduct a comprehensive comparison because there is now a proliferation of such indices (see Maassen, 2000, for a scholarly review of many of these). Moreover, technical arguments abound over even basic issues in this latter field (e.g., see Maassen, 2004; Temkin, 2004; Wise, 2004). These are based mainly on differing interpretations of classical test theory and are beyond the scope of the present paper.

Therefore, attention will be limited to the reliable change index (RCI) proposed by Jacobson and colleagues (e.g., Jacobson & Truax, 1991) because it was the first index so named and remains the most widely used. The original version of this index used the following equation to infer change:

$$\mathrm{RCI} = \frac{x_2 - x_1}{\mathrm{SEM}}, \tag{18}$$

where x_1 and x_2 are the individual's scores at test and retest, respectively, and SEM represents the standard error of measurement of the test and is defined as

$$\mathrm{SEM} = s_x \sqrt{1 - r_{xx}}, \tag{19}$$

where s_x is the standard deviation of the test and r_{xx} is the test–retest reliability coefficient. The standard error of measurement represents the standard deviation of obtained scores around a patient's hypothetical true score. Thus, the original version of the RCI is not fit for the purpose because it uses the standard error of measurement for a single score to test for a reliable difference between two scores. Christensen and Mendoza (1986) therefore suggested that the appropriate standard error was the standard error of measurement of the difference (SE_{diff}) between two scores. Because the same test is used on two occasions, the standard equation for the SE_{diff} was simplified to

$$\mathrm{SE}_{\mathrm{diff}} = \sqrt{2(\mathrm{SEM}^2)}. \tag{20}$$

Jacobson and colleagues (e.g., Jacobson & Truax, 1991) adopted this standard error, and thus the RCI was modified to become

$$\mathrm{RCI} = \frac{x_2 - x_1}{\mathrm{SE}_{\mathrm{diff}}}. \tag{21}$$

This quantity is treated as a standard normal deviate, and if it exceeds 1.96 the patient is considered to have demonstrated reliable change (e.g., it is unlikely that the difference solely reflects measurement error).

From a neuropsychologist's point of view the RCI, either in its original or modified form, is not a very practical proposition because practice effects are ubiquitous in our area of inquiry. If a test has a large practice effect and high reliability, then very large numbers of individuals will exhibit differences that are classified as reliable. However, neuropsychologists are primarily interested in whether any positive change exceeds the practice effect and can thus be attributed to an improvement in the patient's cognitive functioning. Furthermore, if the question at issue is whether the patient's cognitive functioning has deteriorated, practice effects will partially mask any such changes (if the expected practice effect for a test is large, then identical scores on the two occasions would suggest deterioration in functioning, but the RCI is blind to this). In contrast, in the regression approach to quantifying change, practice effects are incorporated into the predicted score against which the patient's score at retest is compared. Note, however, that further variants on the RCI have been proposed that attempt to deal with this problem by subtracting the mean practice effect from the difference between test and retest scores (Chelune et al., 1993; Maassen, 2000).

All of the RC indices discussed have a common feature that is not shared with the regression approach proposed here: They treat the statistics of the sample used in their computation as fixed parameters; that is, it is assumed the standard deviations and reliability coefficients were measured without error. With very large samples, this assumption can be reasonable; that is, the sample statistics provide sufficiently accurate estimates of the parameters. However, the RCI has been widely used with statistics obtained from small to moderate samples; in these circumstances there will be considerable uncertainty as to the true reliability of the measure. In contrast, the regression method proposed here factors in the uncertainty arising from sampling error.

A further difference between the regression approach and the RCI lies in the nature of their null hypotheses. In the case of the RCI, there is one null hypothesis: The difference between the patient's observed scores is due to measurement error; the two scores index the same underlying true score. In the case of the regression approach as set out in the present paper, there is also one null hypothesis: The discrepancy between the patient's predicted and obtained scores is an observation from the population sampled to build the regression equation. However, because different populations can be sampled to answer different questions, the null hypothesis tested is limited only by the ingenuity of the investigator and the availability of suitable equations or suitable data with which to build equations.

For example, suppose a patient is suspected of being in the early stages of Alzheimer's disease and is assessed at an initial interview and at a follow-up period. These scores can be analyzed using a regression equation derived from a healthy sample retested over a similar interval; rejection of the null hypothesis (i.e., the presence of a significant discrepancy) can be used to support the conclusion that the change is not consistent with healthy memory functioning. On the other hand, suppose that the scores of a patient with suspected AD were entered into a regression equation built from an early stage AD sample retested over a similar interval. A significant discrepancy in favor of the actual score at retest raises the possibility of misdiagnosis; that is, the discrepancy is unlikely to have been drawn from the population of discrepancies found in AD.

Furthermore, the method can be used to test for the effects of an intervention. For example, and as previously noted, Chelune et al. (1993) have shown that the baseline and retest scores of patients undergoing surgery for temporal lobe epilepsy (TLE) can be analyzed using an equation built in a TLE sample of patients who had not undergone surgery. In this scenario, a significant discrepancy supports the conclusion that the surgery has affected memory performance. That is, the patient's discrepancy is not an observation from the population of discrepancies found in TLE cases not undergoing surgery (see the worked example in the next section for yet another further form of null hypothesis that also involves TLE).

Finally, demographic variables can exert an influence on change from test to retest; for example, it may be that practice effects on a test diminish with age. These variables can be included as predictor variables in a regression approach in order to gain a more precise predicted retest score (Chelune, 2003; Crawford, 2004; Temkin et al., 1999). This is not possible using the RCI approach to detecting change.

In summary, the principle advantage of the regression-based approach to studying change in the individual case is that it is much more flexible than the RCI and is more in keeping with neuropsychologists' assessment aims. Therefore, we are in agreement with Sawrie's (2002) suggestion that a regression-based approach to inferring change should be the preferred approach. The obvious caveat is that the present regression-based method should be used in preference to the widely used $s_{Y \cdot X}$ method for all the reasons outlined earlier.

Computer Programs Implementing These Methods and an Example

Programs implementing the methods covered in the present paper are available. The program for the bivariate case (regdiscl.exe) prompts the user for the *N* of the sample used to build the equation, *b* (the slope of the regression line), *a* (the intercept), the mean of *X*, and the standard deviations of the *X* and *Y* variables. The user then enters a patient's *X* and *Y* scores. The sample statistics are saved to a file and reloaded when the program is rerun. Therefore, when the program is used with a subsequent patient, the required data for the new patient can be entered, and results obtained, in a few seconds. The program has the option of clearing the sample data to permit entry of a new equation if required.

The output consists of the predicted score; the standardized discrepancy between the predicted and obtained scores; the effect size index for the difference; 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. The program for the vector case (regdisclv.exe) has the same basic features but requires entry of the (lower) diagonal correlation matrix for the *X* variables (the program performs the matrix inversion).

To illustrate the use of the methods and the programs, suppose that a neuropsychologist has a regression equation for predicting memory functioning following a new surgical intervention for temporal lobe epilepsy. Because the technique is new, the equation was built using baseline and retest data from the first 25 patients undergoing the procedure. The mean and standard deviation on the memory test administered at baseline were 90 and 14, respectively, and the mean and standard deviation at retest were 95 and 15, respectively. The slope (*b*) for the equation predicting retest scores from baseline scores was 0.8893, and the intercept (*a*) was 14.964. On the basis of some clinical feature, the neuropsychologist wishes to test the directional hypothesis that a new patient will obtain a poorer than expected outcome.

Entering the sample's statistics into the program for the bivariate case (regdiscl.exe), together with the patient's baseline score (94) and follow-up score (81), reveals that the patient's predicted score is 98.6, and the standardized difference between this score and the obtained score is −2.01. As noted, under the null hypothesis, this standardized difference is distributed as *t* on $N - k - 1 = 23$ *df*. The patient's obtained score is significantly below the score predicted from her or his baseline score ($p = .028$, 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 2.8%, and the 95% confidence limits are from 0.29% to 9.18%. Thus, the patient exhibits a large and statistically significant standardized difference between the obtained and predicted score. The follow-up score is significantly poorer than expected given the patient's baseline, and the discrepancy is estimated to be fairly unusual. Because of the modest sample size, it can be seen that there is considerable uncertainty associated with the estimate of abnormality.

As a further illustration of the methods, suppose that all the statistics were as above, except that the sample size used to build the equation was 200; this could be taken as an example of a standard surgical procedure where a larger series of patients has accumulated. Substituting this *N* into the program reveals that the standardized discrepancy is −2.09. Once again the discrepancy is statistically significant but with a smaller *p* value, $p = .019$ (the power to detect a discrepancy is higher because of the larger sample size). The point estimate of abnormality is also therefore more extreme (1.90%) and, although there is considerably less uncertainty over the abnormality of the discrepancy than in the previous example, the level of uncertainty is still appreciable: The 95% confidence limits for the abnormality of the discrepancy are 0.95% to 3.28%.

Note, however, that in this second example the neuropsychologist can be particularly confident that the patient's discrepancy is unusual. It is unlikely that more than 3.28% of the population would exhibit a discrepancy larger than the patient's. As noted earlier, if the clinician or researcher is interested solely in whether the discrepancy may be more common than the point estimate indicates (i.e., the clinician is unconcerned with whether it may also be even more unusual), it would be appropriate to examine the one-tailed upper limit. In this example, the 95% upper limit on the abnormality of the discrepancy is 3.00%, thereby underlining that the discrepancy is unusually large. With an *N* of 25, the one-sided upper limit is 7.51%, compared to the corresponding two-sided limit of 9.18% reported earlier.

The computer program implementing the method for the bivariate case can be used with existing published regression equations in neuropsychology provided that, as would usually be the case, the basic statistics are available. It is also possible to use the vector program with existing equations, but the requirement of the correlation matrix for the predictor variables imposes limitations. Although the American Psychological Association (2001) strongly recommends that authors reporting multiple regression analyses should include the correlation matrix as a table, this is not always done.

Compiled versions of both programs can be downloaded as zip files from http://dx.doi.org/10.1037/0894-4105.20.3.259.supp or from the following Web site address: http://www.abdn.ac.uk/~psy086/dept/regdisc.htm

Caveats on the Use of These Methods

It should be noted that the validity of inferences made using these methods 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 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 fanlike appearance on a scatterplot), then this assumption would be violated. For those building equations, there are various potential remedies for this and other problems. For example, logarithmic transformations can be applied to the predictor variables (most textbooks on regression cover this topic). When the methods are used with existing published regression equations, it is to be hoped that the peer review process will have provided a degree of quality control so that serious violations of assumptions will be uncommon.

It should also be noted that some regression equations used in neuropsychology will use power functions of the predictor variables (Temkin et al., 1999). For example, in developing regression-based norms, it may be that there is a nonlinear relationship between age and performance on a neuropsychological test. To model this, power functions of age (i.e., age^2) can be incorporated as predictors. The programs described above have no way of knowing if power functions are included in the equations, and therefore the user would have to apply these prior to entering the values for the predictor variables (i.e., in the foregoing example, the square of the individual's age would have to be entered into the program). When used with published equations, a further complication arises in that the summary statistics provided for the sample used to build the equation may not include the mean and standard deviation for age^2 or for the power functions for other variables used in an equation. These summary statistics are required.

Conclusion

The use of regression equations in neuropsychology is widespread but also unusual in that, in the majority of applications, the score on the dependent or criterion variable is known. The neuropsychologist's interest is therefore focused on whether the discrepancy between a patient's predicted and obtained score is abnormal. The treatment of regression in most textbooks is on the standard case in which the score on the criterion variable is unknown. As such, the method of obtaining confidence limits on a predicted score presented in such texts is not particularly helpful in neuropsychological applications.

However, as shown, the standard error used in forming these confidence limits can instead be used to provide a significance test on the discrepancy between a predicted score and a known score (i.e., the obtained score) on the criterion variable. This significance test also simultaneously provides a point estimate of the abnormality of the discrepancy. Confidence limits on the abnormality of this discrepancy were derived in the present paper, and it was shown that these have a wide range of applicability in neuropsychology. Unlike the significance test, which is liable to be of most use when the sample used to build an equation is relatively modest in size (because the technically incorrect but widespread alternative method will give a reasonable approximation when *N* is large), the uncertainty associated with estimates of the abnormality of a discrepancy is appreciable even with very large samples and should therefore be quantified. The method for performing the significance test and for forming confidence limits is complex, particularly in the vector case, but the accompanying computer programs means that neuropsychologists need never perform the computations involved.

References

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

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 28,* 610–611.

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: 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.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17,* 305–308.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Crawford, J. R. (1996). Assessment. In J. G. Beaumont, P. M. Kenealy, & M. J. Rogers (Eds.), *The Blackwell dictionary of neuropsychology* (pp. 108–116). London: Blackwell.

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., & Howell, D. C. (1998a). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12,* 482–486.

Crawford, J. R., & Howell, D. C. (1998b). 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.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

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

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

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

Heaton, R. K., Grant, I., Ryan, L., & Matthews, C. G. (1996). Demographic influences on neuropsychological test performance. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (2nd ed., pp. 141–163). New York: Oxford University Press.

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.

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

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Kennedy, W. J., & Gentle, J. E. (1980). *Statistical computing.* New York: Marcel Dekker.

Knight, R. G., & Shelton, E. J. (1983). Tables for evaluating predicted retest changes in Wechsler Adult Intelligence Scale scores. *British Journal of Clinical Psychology, 22,* 77–81.

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.

Maassen, G. H. (2000). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology, 22,* 622–632.

Maassen, G. H. (2004). The standard error in the Jacobson and Traux Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society, 10,* 888–893.

McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Lüders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist, 7,* 300–312.

Paolo, A. M., Ryan, J. J., Troster, A. I., & Hilmer, C. D. (1996). Demographically based regression equations to estimate WAIS–R subtest scaled scores. *Clinical Neuropsychologist, 10,* 130–140.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal.* Cambridge, England: Cambridge University Press.

Sawrie, S. M. (2002). Analysis of cognitive change: A commentary on Keith et al. (2002). *Neuropsychology, 16,* 429–431.

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.

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

Temkin, N. R. (2004). The standard error in the Jacobson and Traux Reliable Change Index: The "classical approach" leads to poor estimates. *Journal of the International Neuropsychological Society, 10,* 899–901.

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.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82,* 50–59.

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

Zar, J. H. (1996). *Biostatistical analysis* (3rd ed.). London: Prentice Hall.

Received September 2, 2005
Revision received October 21, 2005
Accepted October 28, 2005 ■