Journal of the International Neuropsychological Society (1999), **5**, 357–369.
Copyright © 1999 INS. Published by Cambridge University Press. Printed in the USA.

Detecting significant change in neuropsychological test performance: A comparison of four models

NANCY R. TEMKIN,[1,2] ROBERT K. HEATON,[5] IGOR GRANT,[5,6] AND SUREYYA S. DIKMEN[1,3,4]

Departments of Neurological Surgery,[1] Biostatistics,[2] Rehabilitation Medicine,[3] and Psychiatry and Behavioral Sciences,[4]
University of Washington, Seattle, WA; Department of Psychiatry,[5] University of California, San Diego, La Jolla, CA,
and VA San Diego Healthcare System,[6] San Diego, CA

(Received October 13, 1997; Revised June 18, 1998; Accepted July 13, 1998)

Abstract

A major use of neuropsychological assessment is to measure changes in functioning over time; that is, to determine whether a difference in test performance indicates a *real* change in the individual or just chance variation. Using 7 illustrative test measures and retest data from 384 neurologically stable adults, this paper compares different methods of predicting retest scores, and of determining whether observed changes in performance are unusual. The methods include the Reliable Change Index, with and without correction for practice effect, and models based upon simple and multiple regression. For all test variables, the most powerful predictor of follow-up performance was initial performance. Adding demographic variables and overall neuropsychological competence at baseline significantly but slightly improved prediction of all follow-up scores. The simple Reliable Change Index without correction for practice performed least well, with high error rates and large prediction intervals (confidence intervals). Overall prediction accuracy was similar for the other three methods; however, different models produce large differences in predicted scores for some individuals, especially those with extremes of initial test performance, overall competency, or demographics. All 5 measures from the Halstead–Reitan Battery had residual (observed − predicted score) variability that increased with poorer initial performance. Two variables showed significant nonnormality in the distribution of residuals. For accurate prediction with smallest prediction–confidence intervals, we recommend multiple regression models with attention to differential variability and nonnormality of residuals. (*JINS*, 1999, *5*, 357–369.)

Keywords: Reliability, Stability, Prediction intervals, Confidence intervals, Test–retest data

INTRODUCTION

One of the major uses of neuropsychological assessment is to measure changes in functioning over time. Such changes, on the positive side, may include improved functioning due to treatments or to spontaneous recovery following brain injury or toxic exposure. It is also important to be able to detect deteriorating functioning due to disease progression, treatment side effects, or other new brain insults. While statistical methods for comparing groups on neuropsychological change are relatively straightforward, techniques for detecting significant or "real" neuropsychological changes in *individuals* are less well developed.

Repeated administrations of neuropsychological tests frequently yield varying results, even in people who have not experienced any true change in neurobehavioral status. This is due to less than perfect reliability of test instruments, as well as practice effects, fluctuations in test taking attitudes, and other factors. Moreover, these influences are or may be different for different tests, and for different types of people, and many questions exist about how much of a change in test score is significant or unusual for a particular individual and test. More fundamentally, it is unclear what statistical approach is best suited for predicting a follow-up neuropsychological test score and what factors should be considered in the prediction model. Potentially important factors include baseline performance on the test in question, various participant variables (age, history of risks for

Dr. Erin D. Bigler served as action editor during the course of this review.

Reprint requests to: Nancy R. Temkin, Department of Neurological Surgery (Box 359924), University of Washington, Seattle, WA 98104-2499. E-mail: temkin@biostat.washington.edu

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

neuropsychological impairment, general level of neuropsychological competence at baseline), features of the test in question (test–retest reliability, practice effects, floor–ceiling on test scores), duration of the test–retest interval, and indicators of the participant's state at the time of the testings (mood, cooperativeness, medications consumed).

Furthermore, after settling on a method for predicting a follow-up neuropsychological test score, additional questions remain about how best to establish the interval around the prediction that contains commonly observed differences between the predicted values and the actual scores on repeat testing when there has been no real change in the individual. We will call these intervals *prediction intervals* (Kleinbaum & Kupper, 1978) to be consistent with the statistical literature, although they also have been called *confidence intervals* (McSweeny et al., 1993; Sawrie et al., 1996). In this paper, we will use intervals that are expected to contain 90% of the differences between actual and predicted test scores. Discrepancies outside the prediction interval are considered to indicate "significant" change. Usually one tries to define intervals so that, in a population that is stable, one would have 5% of cases show "significant deterioration" and 5% show "significant improvement." But how should these intervals be determined? For example, can they be based on the normal distribution? Might different prediction intervals be needed for different types of people (old *vs.* young, generally high functioning *vs.* low functioning)?

In the present article we will address several of these questions using test–retest data from a large sample of neurologically stable adults. More specifically, we will compare four models for assessing change in selected neuropsychological variables, using multiple measures of prediction accuracy: standard deviation of the differences between observed and predicted scores; deviations from expected percentages of participants who score outside of designated prediction intervals in both the positive and negative directions; and the length of the prediction interval, that is, the minimum positive and negative deviations from predicted scores that define significant change (given comparable error rates, a shorter length indicates more sensitivity to detect true change). For each model we also present methods for calculating *z* scores to facilitate comparing the direction and degree of change across measures.

The first, and simplest model to be considered here is the Jacobson and Truax (1991) Reliable Change Index (RCI). This method bases the significance of a change in any individual test score on the difference between the initial and retest scores for the normative subject sample. If the absolute value of this change exceeds the standard deviation of the test–retest differences in the norming sample, multiplied by the *z*-score cut point that defines a designated percentile in the normal distribution, the change is considered reliable (i.e., unlikely to occur by chance). The most commonly used cutoff percentage point is the 95th, $Z_\alpha = Z_{.95} = 1.645$; this defines a prediction interval that should include 90% of individuals like those in the norming sample, with 5% outside the interval on the lower end and 5% outside on the upper end. In the terms defined above, the RCI model is equivalent to the predicted retest value equaling the initial score and the prediction interval extending in each direction 1.645 standard deviations of the test–retest differences.

A second model attempts to improve upon the RCI by including an adjustment for practice effects. In this model, the predicted retest score is the person's baseline score plus the mean practice effect for the normative sample, and the procedure for defining an unusual deviation from the predicted value is the same as in the basic RCI procedure (Chelune et al., 1993).

The third model uses linear regression of the retest scores on the initial scores in the norming sample to generate a formula for predicting a follow-up score from any baseline score (McSweeny et al., 1993). This approach provides correction for both practice effects and regression toward the mean. The prediction interval extends in each direction 1.645 times the standard deviations of the residuals estimated by the regression. That is, a retest score is considered unusual if the difference between it and its predicted value exceeds the residual standard deviation from the norming sample times 1.645, the *z*-score cutoff for a 90% interval based on the normal distribution.

The last model to be considered here uses stepwise linear regression for the prediction of retest scores on the basis of multiple factors that could be important. In addition to the baseline scores on the test in question, variables considered in this prediction model were test–retest interval, demographic variables, and a measure of overall neuropsychological competence at baseline. Furthermore, the possibility of a nonlinear relationship between initial and retest scores was considered by including the square and the cube of the initial score in the variable selection. Similarly, the square and cube of the test–retest interval were evaluated for inclusion. For the fourth model, one can use the analogous method for determining an unusual deviation from the predicted retest score as in the simple linear regression approach (Model 3), with the residual standard deviation being obtained from the multiple regression.

All of the methods discussed so far are based on the assumption that the residuals follow a normal distribution. This assumption is sometimes false. Thus we examine the distribution of residuals and, for several measures where the residuals have a decidedly nonnormal distribution, we also present distribution-free intervals.

Finally, to explore whether accuracy of prediction is constant across different levels of predictor variables, we divided cases into subgroups based on their predictor values and calculated the standard deviation of the residuals for each subgroup. If these standard deviations were substantially higher for one subgroup than another (e.g., for older than for younger participants), this implies that different subgroups need different cutoffs for defining unusual deviations from their predicted retest values. Although this extra step could be taken for any of the prediction models, the

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

relevant analyses were performed only for the multiple regression model (Model 4) for illustrative purposes.

METHODS

Research Participants

The participants were 384 normal or neurologically stable individuals who were tested twice as part of several longitudinal studies. All were at least 15 years of age. One hundred thirty-eight participants had no recent trauma history and were friends of head-injured cases; these *friend controls* had a scheduled test–retest interval of 11 months. One hundred twenty-one had suffered a recent traumatic injury that spared the head; these we call *trauma controls*. They were tested for baseline 1 month after trauma, and then 11 months later. All of the friend controls and trauma controls were tested at the University of Washington under the direction of one of us (S.S.D.). Twenty percent of friend controls and 46% of trauma controls had preexisting conditions that might affect test performance, the most common being alcoholism or a head injury in the past. The remaining participants in these groups denied any history of conditions that might be expected to affect brain function. The final 125 participants were enrolled in longitudinal research projects at multiple sites under the supervision of the neuropsychology laboratories at the University of Colorado (R.K.H.) or the University of California at San Diego (I.G.); these individuals had no history of trauma or disease involving the brain. The scheduled test–retest intervals of these participants ranged from approximately 2 to 12 months. These samples were chosen to represent a range of demographics pertinent to neuropsychological status in neurologically stable individuals.

Test Measures

The Halstead–Reitan Neuropsychological Test Battery (HRB) and the Wechsler Adult Intelligence Scale (WAIS) were administered to all participants according to instructions contained in their respective manuals (Reitan & Wolfson, 1993; Wechsler, 1955) For the present analyses, the following representative measures were chosen from these batteries: the WAIS Verbal and Performance IQs, the Category Test (number of errors), the Tactual Performance Test (TPT)–Total Time (minutes per block for the combined trials with dominant, nondominant, and both hands), Trails B (number of seconds to complete), Halstead Impairment Index, and Average Impairment Rating (AIR; Russell et al., 1970). In order to reduce testing time and patient fatigue, time limits were imposed on the Trail Making Test (Trails B = 300 s) and the Tactual Performance Test (10 min each for trials with dominant, nondominant, and both hands).

Data Analyses

The data analytic approaches for Models 1 through 3 are outlined above in the Introduction. For the multiple regression model (Model 4), variables evaluated as potential predictors for Time 2 score on each neuropsychological measure include the score on that measure at Time 1, its square and cube (to allow for a nonlinear relationship), the overall neurobehavioral competence at Time 1 as estimated by the Average Impairment Rating, the test–retest interval (in months), demographic information, and presence or absence of preexisting conditions that could affect brain function. (The latter were prior hospitalization for head trauma or treatment for alcoholism, each coded 0 if absent, 1 if present.) Demographic information included age (in years), years of education (counted as 12 for a student currently in high school), an indicator for current high school student status (0 if not in high school at first testing, 1 if in high school), sex (0 if male, 1 if female), and race (0 if White, 1 if Nonwhite). A variation that omits the measure of overall neurobehavioral competence allows one to determine how much demographic information, nonlinear terms, and interval alone add to the prediction by Time 1 score.

Stepwise linear regression was used to predict retest values based upon the measures just described. At each step, a variable was added to the prediction model if it had the highest partial correlation among variables not already in the equation, and if the significance level associated with the variable was under .05; conversely, a variable was removed from the model if its significance level rose above .10 as other variables were added.

It should be noted that, with both the bivariate and multivariate regression models (Models 3 and 4), it is possible for a participant to have no change or even slight worsening in their score on a retest, but because of adjustment for regression to the mean or other predictors, to have the equations indicate that this represents improvement over what is predicted. This occurred very rarely, however, and for the present demonstrations we conservatively elected to count this as *no change* rather than *improvement*. We treated similarly those cases where no change or an actual improvement from first to second testings was indicated by the equations to represent deterioration, although we acknowledge that true deterioration could be manifest by the absence or diminution of an expected practice effect. In the over 2,500 predictions evaluated for this paper, these two situations arose only seven times.

As was pointed out above, in order to identify the “usual” ranges of retest scores, we examined residuals from the prediction models (i.e., the differences between actual and predicted retest scores). Most commonly, this is done making the assumption that the residuals follow a normal distribution. However, it is possible that, for some measures, the residuals have a skewed distribution, so that the predicted range based on the normal distribution actually shows a higher or lower than desired percent of the normative sample being classified as “unusual” in one direction. To evaluate these possibilities, we examined the actual percent of cases below the theoretical 2.5, 5, 10, 20, 80, 90, 95, and 97.5 percentage points based on the normal distribution. If two or more of the actual percentages differed significantly

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

from the normal-distribution-based values at the testwise .05 level using a binomial test, we considered the assumption of normality to be violated and calculated additional intervals around predicted retest scores based on the observed distribution of the residuals; that is, for these variables, we will present the observed 5th and 95th percentiles of the residuals as "distribution-free" 90% prediction intervals.

To examine whether the accuracy of the prediction is constant across different levels of the predictor variables, we divided the cases into subgroups based on their predictor values and calculated the standard deviation of the residuals for each subgroup. If the standard deviation was at least 25% higher in one subgroup than another, we calculated different prediction intervals for different subgroups. As noted above, this could be done for any of the methods but is provided for the multiple regression method for illustrative purposes.

For any of the methods, one can transform the retest scores into standardized z scores by calculating z score $=$ sign $\times$ (observed retest score $-$ predicted retest score)/residual standard deviation. Sign equals $+1$ for measures where a higher score indicates better performance (VIQ, PIQ) and equals -1 for measures where a lower score indicates better performance (Category, TPT total, Trails B, Halstead Index, AIR). Thus a positive z score indicates better than predicted retest performances.

RESULTS

The subject sample's demographic characteristics and test–retest intervals are described in Table 1. Although only 14% of the sample is over age 54 and only 11% is Nonwhite, both low and high education levels are well represented. All test–retest intervals are between 2 and 16 months, and within that range, all intervals except the longest contain at least 50 individuals.

Table 1. Demographic information for the participant sample ($N = 384$)

Variable	*N* (%)	*M* (*SD*)
Age (years)		
15–24	137 (36)	34.2 (16.7)
25–34	105 (27)	
35–44	61 (16)	
45–54	25 (6)	
55–64	19 (4)	
65–74	26 (7)	
75 or older	11 (3)	
Sex		
Male	253 (66)	
Female	131 (34)	
Racial–ethnic category		
White (including Hispanic)	340 (89)	
Nonwhite	44 (11)	
Education		12.3 (2.7)*
High school student	38 (10)	
Less than high school	89 (23)	
High school graduate	107 (28)	
Some college	112 (29)	
College graduate	38 (10)	
Test–retest interval (months)		9.1 (3.0)
2.3– 4.9	55 (14)	
5.0– 7.9	56 (15)	
8.0–10.9	166 (43)	
11.0–13.9	106 (28)	
15.8	1 (.03)	
Preexisting alcoholism (%)	57 (15)	
Prior head injury (%)	27 (7)	

*Excluding current high school students

Predicting Retest Scores

The means and standard deviations of the scores at each testing, as well as of the Time 2 minus Time 1 difference, are given in Table 2. The ranges of neuropsychological scores at the initial testing are also given. Inspection of the mean Time 2 minus Time 1 difference scores reveals considerable variability in the amounts of practice effects they show. As defined by change from Time 1 to Time 2 compared to the standard deviation at Time 1, the Category Test and PIQ show relatively large practice effects whereas the VIQ and Trails B show relatively small practice effects.

Prediction results based on regression (Models 3 and 4) are given in Table 3. This table gives the test–retest correlations and the standard deviation of the residuals, as well as the slope (unstandardized beta) and intercept (constant) needed to predict the retest score based only upon the initial test score (Model 3). The table also includes the multiple correlation after all related predictors were included, as well as the standard deviation of the residuals after all predictor variables were entered (Model 4). The additional predictors also are shown in the order they entered. Except for Halstead's Impairment Index, the initial test result always entered the prediction model first. All of these predictions improved significantly with the addition of variables other than initial test score, although the magnitude of the increased correlation and reduced standard deviation of residuals usually was not large. This is true particularly for the prediction of VIQ2: After VIQ1 was considered, education, age, and test–retest interval improved prediction a statistically significant amount; nevertheless, the improvement in correlation was slight, in the 3rd decimal place. AIR, the measure of overall neuropsychological competence, entered most predictions either second or first. That measure is based on a whole battery of tests. As seen from the italicized entries in Table 3, predictions that are almost as good can be obtained using only demographics and squares and cubes of the Time 1 score if AIR is not available.

The prediction equations for the multiple regression models (Model 4 and its variant) are given in Table 4. By substituting values for an individual of interest, one can calculate the predicted value at the second testing. For example, if a 60-year old scored .20 min per block on TPT Total (an un-

Table 2. Summary of initial, retest, and test–retest difference scores

	Time 1				Time 2		Difference ($T2 - T1$)	
Measure	*M*	(*SD*)	Minimum	Maximum	*M*	(*SD*)	*M*	(*SD*)
VIQ	108.4	(13.7)	69	149	109.5	(14.0)	1.1	(4.8)
PIQ	108.5	(11.5)	73	135	113.6	(12.7)	5.1	(6.4)
Category	41.0	(26.1)	4	145	30.4	(25.0)	−10.5	(14.1)
TPT Total	.52	(0.49)	.16	6.0	.43	(0.33)	−.09	(0.29)
Trails B	72.0	(45.2)	25	276	68.2	(46.1)	−3.9	(21.6)
Halstead Index	.29	(.28)	0.0	1.0	.24	(.27)	−.05	(0.17)
AIR	1.02	(.56)	.17	3.36	.88	(.55)	−.14	(0.22)

usually good score) and 1.00 on Average Impairment Rating at the first testing, their TPT Total score at a second testing is predicted to be $\text{TPTtotal2} = .590 \times \text{TPTtotal1} - .0340 \times \text{TPTtotal1}^2 + .0844 \times \text{AIR1} + .00200 \times \text{age} - .0155 = .590 \times .20 - .0340 \times .20^2 + .0844 \times 1.00 + .00200 \times 60 - .0155 = .29$. Using only the Time 1 score as a predictor, this person would have a predicted score of $(.55) \times (.20) + .14 = .25$; using the RCI with practice, the predicted score would be .11. The difference in values predicted by the models is substantial. For a more usual case, such as a 20-year-old with Time 1 scores of .60 on TPT Total and 1.50 on AIR, the predictions are much closer: .51 by Model 2, .47 by Model 3, and .49 by Model 4.

Prediction equations for the Average Impairment Rating and Trails B are portrayed graphically in Figure 1 for a few predictor values. Figure 1A shows the predicted Time 2 score

Table 3. Summary of the regressions. The slope and intercept are given for the prediction based on only initial scores as is the multiple correlation of retest score with initial score and with all predictors entering. The residual standard error after the initial score and after all selected variables were entered are also shown. The specific predictors for each measure are shown in the order of predictor entry. For models where AIR1 entered as a measure of overall competence, the results from the variation that excluded that predictor are shown below in italics.

	Model 3				Model 4		
Measure	Correlation with initial	Residual *SD* after initial	Slope	Intercept	Correlation with all	Residual *SD* after all	Multiple regression predictors[1]
VIQ	.94	4.8	.95	6.1	.94	4.7	VIQ1, Ed, Age, Interval
PIQ	.86	6.4	.95	10.7	.88	6.1	PIQ1, AIR1, Interval, Ed, Race
					.87	*6.2*	*PIQ1, Ed, Race, current high school student*
Category	.84	13.3	.80	−2.6	.88	11.9	Category1, AIR1, Race, Age, Ed
					.87	*12.3*	*Category1, Age, Race, Ed, Category1 squared*
TPT Total	.88	.15	.55	.14	.91	.13	TPTtotal1, AIR1, TPTtotal1 squared, Age
					.91	*.13*	*TPTtotal1, TPTtotal1 squared, Age, TPTtotal1 cubed, Race*
Trails B	.88	21.3	.90	3.5	.90	19.6	Trails B1, AIR1, Age, TrailsB1 squared, Ed, Prior head injury
					.90	*19.9*	*Trails B1, Age, TrailsB1 squared, Ed, Prior head injury, Race*
Halstead Index	.82	.16	.81	.01	.87	.14	AIR1, HI1, Age, HI1 squared, Race, Sex, HI1 cubed
					.85	*.14*	*HI1, Age, HI1 squared, Race, Ed*
Average Impairment Rating	.92	.21	.90	−.04	.94	.19	AIR1, Age, Race, Ed, AIR1 squared, AIR1 cubed, Interval

[1] Initial test scores are indicated by the test name followed by a 1.
Note. AIR = Average Impairment Rating, Ed = education, HI = Halstead Index, Interval = test–retest interval.

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

Table 4. Equations to predict retest score based on initial test result and other variables. Equations in italics exclude AIR as a potential predictor.

$\mathrm{VIQ2} = .901 \times \mathrm{VIQ1} + .0372 \times \mathrm{age} + .464 \times \mathrm{education} - .194 \times \mathrm{interval} + 6.58$

$\mathrm{PIQ2} = .803 \times \mathrm{PIQ1} - 3.34 \times \mathrm{AIR1} - 2.38 \times \mathrm{race} + .347 \times \mathrm{education} - .269 \times \mathrm{interval} + 28.5$

$\mathit{PIQ2} = .879 \times \mathit{PIQ1} + .607 \times \mathit{education} - 3.10 \times \mathit{race} + 2.59 \times \mathit{current\ high\ school\ student} + 11.5$

$\mathrm{Category2} = .520 \times \mathrm{category1} + 10.8 \times \mathrm{AIR1} + 7.86 \times \mathrm{race} + .143 \times \mathrm{age} - .780 \times \mathrm{education} + 1.77$

$\mathit{Category2} = .430 \times \mathit{category1} + .00214 \times \mathit{category1}^2 + .256 \times \mathit{age} + 10.0 \times \mathit{race} - 1.11 \times \mathit{education} + 1.58$

$\mathrm{TPTtotal2} = .590 \times \mathrm{TPTtotal1} - .0340 \times \mathrm{TPTtotal1}^2 + .0844 \times \mathrm{AIR1} + .00200 \times \mathrm{age} - .0155$

$\mathit{TPTtotal2} = .820 \times \mathit{TPTtotal1} - .135 \times \mathit{TPTtotal1}^2 + .0118 \times \mathit{TPTtotal1}^3 + .00303 \times \mathit{age} + .0514 \times \mathit{race} - .103$

$\mathrm{TrailsB2} = .355 \times \mathrm{TrailsB1} + .00112 \times \mathrm{TrailsB1}^2 + 14.6 \times \mathrm{AIR1} + .331 \times \mathrm{age} - 1.18 \times \mathrm{education} + 8.39 \times \mathrm{prior\ head\ injury} + 22.1$

$\mathit{TrailsB2} = .518 \times \mathit{TrailsB1} + .000933 \times \mathit{TrailsB1}^2 + .472 \times \mathit{age} - 1.54 \times \mathit{education} + 9.45 \times \mathit{prior\ head\ injury} + 6.96 \times \mathit{race} + 9.31$

$\mathrm{HI2} = -.305 \times \mathrm{HI1} + 1.48 \times \mathrm{HI1}^2 - .908 \times \mathrm{HI1}^3 + .214 \times \mathrm{AIR1} + .00281 \times \mathrm{age} + .0575 \times \mathrm{race} + .0396 \times \mathrm{sex} - .146$

$\mathit{HI2} = .303 \times \mathit{HI1} + .385 \times \mathit{HI1}^2 + .00372 \times \mathit{age} + .0867 \times \mathit{race} - .0078 \times \mathit{education} - .036$

$\mathrm{AIR2} = .188 \times \mathrm{AIR1} + .381 \times \mathrm{AIR1}^2 - .0685 \times \mathrm{AIR1}^3 + .00506 \times \mathrm{age} - .0119 \times \mathrm{education} + .149 \times \mathrm{race} + .00777 \times \mathrm{interval} + .215$

Note. Initial test scores are indicated by the test name followed by a 1, retest scores are indicated by the test name followed by a 2. AIR = Average Impairment Rating, HI = Halstead Index, Interval = test–retest interval in months.

on Average Impairment Rating for any Time 1 score for a White person with a high-school education and a 12-month test–retest interval. Note that the accuracy of the prediction and the size of the prediction intervals are not reflected in Figure 1; these important issues will be addressed later. In Figure 1A, three curves are shown representing participants at different ages: 20, 40, and 70 years old. One can see the curvature in the prediction line and that, even for a fixed initial score, older people tend to score more poorly at retest. For comparison, three straight lines corresponding to expected performance according to Models 1, 2, and 3 also are plotted on each panel. The simplest, shown by dots, predicts the Time 2 score to be identical to that at the first testing, according to the Reliable Change Index. The parallel line below it takes into account the average practice effect shown in Table 2, as would be done in Model 2. Compared to regression-based predictions, both of these prediction procedures tend to estimate values that are more extreme for people with very good or very bad initial scores. The less sloped line, shown by long dashes, plots the regression-based prediction using initial score only; the slope and intercept for this line are given in Table 3.

Figure 1B depicts the Time 2 Trails B score as a function of the Time 1 score and initial score on Average Impairment Rating. This is for a 40-year-old with a high school education and no prior history of head injury or alcohol abuse. For Model 4, one can see that initial overall competence, as estimated by Average Impairment Rating, has a substantial effect on the predicted Trails B retest score even after accounting for initial Trails B score. Again, the three straight lines corresponding to Models 1, 2, and 3 are shown for comparison.

Comparing Prediction Intervals Based on the Four Procedures

Table 5 shows, for each of the four prediction models, the prediction intervals based on the normal distribution as well as the percent of the norming sample outside those intervals in each direction. All of the methods considered are designed so that 5% of the cases should be outside the prediction interval in the direction indicating improvement and 5% should be out in the direction indicating deterioration. As seen in Table 5, for the simplest method based on the Reliable Change Index (in which the predicted value at Time 2 is the Time 1 score) most of the neuropsychological variables have a significantly higher than expected percentage of participants classified as improved and a significantly lower percentage classified as deteriorated. Most of these measures have a practice effect, and a method that ignores it usually suggests that far too many people improve. By accounting for the average practice effect, Model 2 yields percentages much closer to those expected. That does not mean, however, that the differences between the observed and predicted scores follow a normal distribution. In fact, Category Test, TPT Total time, Trails B time, and Halstead Index all have at least two of the checked percentiles significantly different from those based on the normal distribution. Additionally, since Model 2 ignores regression to the mean, as seen in Figure 1, it is likely to predict scores that are too good for those initially scoring well and too bad for those initially scoring poorly. The regression methods take this into account, yielding slightly narrower prediction intervals; for TPT–Total time per block, the prediction interval narrows substantially by using a Model 3 prediction based only on the Time 1 score. Category, TPT Total, Trails B, and Halstead Index show deviations from normality in the residuals from Model 3. Taking other potential predictors into account (in Model 4) narrows the prediction interval by a slight but statistically significant amount for most of the measures considered. TPT time again shows a noticeable decrease in the width of the prediction interval.

Distribution-free intervals

Both Trails B and TPT Total time per block have residuals that are substantially nonnormal for all of the models. Both

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

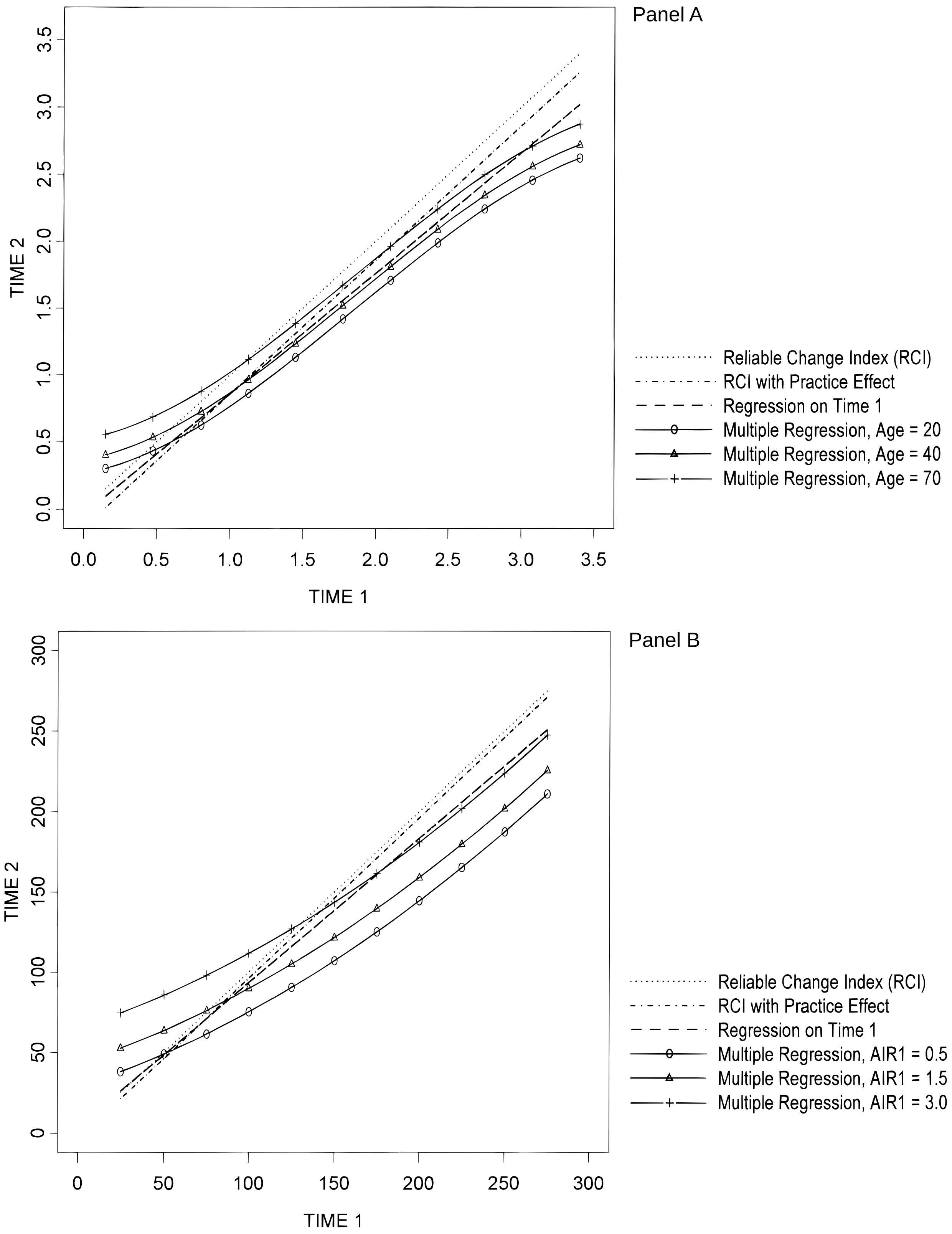

Fig. 1. Time 2 score predicted by the four models as a function of Time 1 score. Panel A presents Average Impairment Rating; the multiple regression predictor (Method 4) is shown for a White person with a high-school education and a 12-month test–retest interval and a selection of ages. Panel B presents Trails–B; the multiple regression predictor (Method 4) is shown for a 40-year-old with a high school education and no prior head injury and a selection of initial scores on Average Impairment Rating (AIR).

Table 5. Normal-distribution-based prediction intervals and the percentages of participants classified as “unusual” by four models for predicting neuropsychological retest scores

	Model 1 Reliable Change Index (RCI)			Model 2 RCI with practice effect			Model 3 Regression on Time 1 score			Model 4 Regression with all predictors		
Measure	Prediction interval[a]	Percent improved	Percent deteriorated	Prediction interval[a]	Percent improved	Percent deteriorated	Prediction interval[a]	Percent improved	Percent deteriorated	Prediction interval[a]	Percent improved	Percent deteriorated
VIQ	±7.9	10*	5	±7.9	3	7	±7.9	4	7	±7.7	4	6
PIQ	±10.6	20*	1*	±10.6	7	3	±10.6	6	3	±10.0	4	4
Category	±23.2	17*	2*	±23.2	6	2*	±21.9	5	5	±19.6	4	4
TPT–Total	±.48	4	1*	±.48	3	1*	±.31	1*	4	±.21	2*	4
Trails–B	±35.6	7	3	±35.6	4	4	±35.1	4	4	±32.3	2*	5
Halstead Index	±.28	14*	4	±.28	3	4	±.26	3*	8*	±.22	5	7
AIR	±.36	14*	0*	±.36	4	5	±.35	5	6	±.32	4	6

[a] Prediction interval indicates the values around the model-predicted Time 2 score that would be expected to be seen in 90% of the norming sample. “Improved” and “Deteriorated” indicate the percentage of participants in the norming sample who were actually classified as unusually better or worse, respectively, at Time 2 based on the indicated method of obtaining the predicted value and a prediction interval based on the normal distribution. Those that differ significantly from the expected 5% are marked with *.

have some skewness as well as outliers that inflate the standard deviation of the residuals. Table 6 gives the distribution-free prediction intervals for these two measures. Because of the outliers, the distribution-free intervals are narrower than the normal-based intervals shown in Table 5 and still result in 5% of these stable individuals classified as each of improved and deteriorated.

Factors affecting variability of the Time 2 scores

Table 7 lists the standard deviation of the residuals from the multiple regression model, subdivided by age, education, initial score, and preexisting conditions. The *poor*, *average*, and *good* ranges for initial score are based on the raw (demographically uncorrected) scaled scores for the norms derived by Heaton et al. (1991) and represent scaled scores less than or equal to 7, between 8 and 11, and greater than or equal to 12. These ranges contain approximately the worst 20%, middle 50%, and the best 30% of Heaton’s normative population. With the notable exception of the WAIS IQs, most measures show substantial differences in variability among groups defined by initial score, and some by several other variables as well. In all cases in which the variability differs substantially, the direction is that poor initial performance or factors associated with poor performance are associated with increased variability on retest. This has major implications for prediction intervals. The intervals given so far will tend to be too wide, hence missing true change, in individuals with good initial performance, and will be too narrow, hence calling normal variability changed performance, in those with poor performance. Table 8 shows revised prediction intervals for Model 4 taking initial performance into account. The number of categories and the values grouped together were chosen based on visual inspection of the scatterplots of residuals by Time 1 score. When the variability of the residuals was calculated for subgroups based on the other predictors shown in Table 7, the results showed much smaller, though in some cases still significant, differences. Thus, although only initial score is explicitly taken into account, these intervals correct to some extent for the other differences in variability, as well. Note that the full range of IQs is presented together because, for these measures, the variability of the retest score did not differ appreciably depending on initial score.

Example

Figure 2 shows the predicted retest scores (horizontal lines) and prediction intervals (bars around the horizontal lines) for TPT Total for the 2 hypothetical individuals discussed earlier. Panel A represents the predictions of a “usual” case,

Table 6. Distribution-free prediction intervals for two measures with frequently nonnormal residuals

	Model 1 RCI		Model 2 RCI with practice effect		Model 3 Regression on Time 1 score		Model 4 Regression with all predictors	
	Prediction interval		Prediction interval		Prediction interval		Prediction interval	
Measure	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
TPT–Total	−.35	.13	−.26	.22	−.14	.25	−.14	.17
Trails–B	−38.5	28.5	−34.6	32.4	−31.4	32.6	−25.4	31.7

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

Table 7. Standard deviations of the residuals from the multiple regression model for subgroups based upon initial levels of predictor variables

Factor	VIQ	PIQ	Category	TPT–Total	Trails–B	Halstead Index	Average Impairment Rating
Time 1 score							
Poor	4.4	5.4	17.0	.26	37.0	.18	.22
Average	4.8	6.4	11.6	.09	15.4	.16	.20
Good	4.7	5.8	4.8	.05	9.8	.09	.14
Age							
≥60	4.8	5.8	15.6	.29	32.0	.18	.19
<60	4.6	6.0	11.3	.09	17.4	.13	.23
Education							
In high school	4.3	6.9	14.7	.05	9.5	.07	.20
<12 years	4.8	5.8	15.4	.16	26.5	.16	.22
≥12 years	4.7	6.0	9.9	.12	17.8	.13	.18
Preexisting alcohol							
Yes	4.6	6.9	12.8	.12	22.8	.18	.21
No	4.7	5.9	11.7	.13	19.0	.13	.19
Prior head injury							
Yes	4.1	5.9	9.9	.15	23.6	.15	.22
No	4.7	6.0	12.0	.13	19.2	.13	.19

with only small variation in the predicted values but sizable differences in the prediction intervals. Panel B represents a more atypical individual (older person with excellent TPT Total initial value), where the different models yield substantially different intervals and conclusions.

Standardized z *scores*

Standardized z scores corresponding to each of the methods discussed can readily be calculated from the information presented. Z scores provide a consistent metric for the measures and, if the residuals are normally distributed with constant variability, allow one to easily calculate the probability of differences this extreme. In general, z score $=$ sign $\times$ (observed retest score $-$ predicted retest score)/residual standard deviation, with sign being $+1$ or -1 depending on whether higher or lower scores indicate better performance. For the first model (Reliable Change Index), the predicted score is the score at initial testing. For the second model (Reliable Change Index with Practice Effect), the predicted score is the score at initial testing plus the average $T_2 - T_1$ difference given in Table 2. For both models, the residual standard deviation is the standard deviation of the $T_2 - T_1$ difference given in the last column of Table 2. For Model 3 (Regression on Time 1 Score), both the prediction equation (slope $\times$ Time 1 score $+$ intercept) and residual standard deviation (residual standard deviation after initial) are given in Table 3. For Model 4 (Regression on All Predictors), the predicted value is obtained using the equations in Table 4. The residual standard deviations are given in Table 7 as a function of the Time 1 score. Note that

Table 8. Prediction intervals based on regression on all selected variables allowing interval width to vary depending on the initial score. The standard deviation of the residuals and whether the intervals are based on the normal distribution (N) or are distribution-free (DF) are also shown.

		90% prediction interval		
Measure—Time 1 score	*SD*	Lower endpoint	Upper endpoint	Basis
VIQ				
Full range	4.7	−7.7	7.7	N
PIQ				
Full range	6.1	−10.0	10.0	N
Category errors				
25 or less	5.2	−8.6	8.6	N
26–59	12.6	−20.6	20.6	N
60 or more	17.0	−28.0	28.0	N
TPT–Total time				
.4 or less	.06	−.10	.10	N
.41–1.24	.14	–.22	.22	N
1.25 or more	.42	−.69	.69	N
Trails–B time				
40 or less	8.6	−14.1	14.1	N
41–99	14.0	−23.0	23.0	N
100 or more	40.4	−66.5	66.5	N
Halstead Index				
.1 or less	.09	−.11	.20	DF
.2 or more	.16	−.27	.27	N
Average Impairment Rating				
.75 or less	.16	−.26	.26	N
.76 or more	.21	−.35	.35	N

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

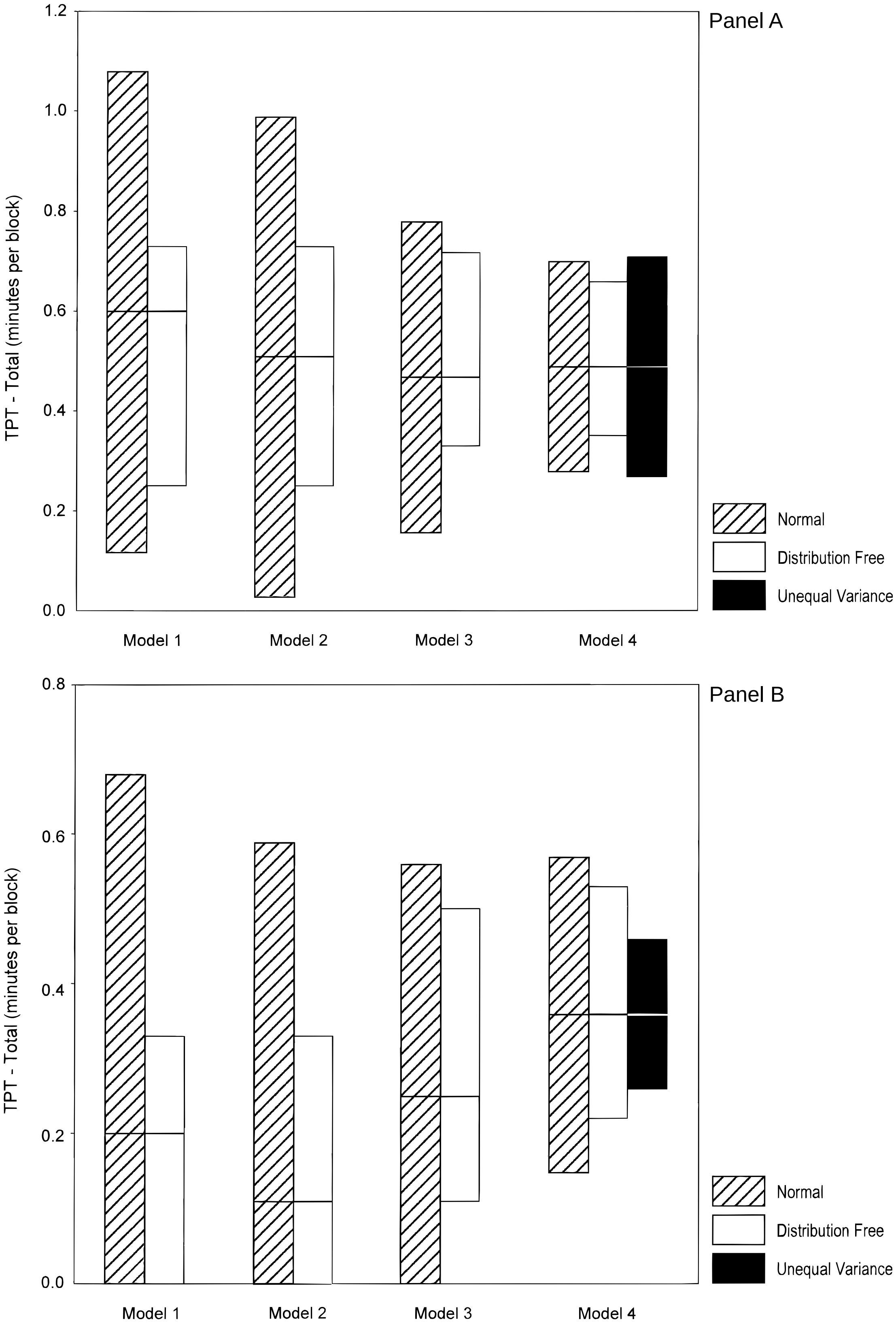

Fig. 2. Examples of prediction intervals for TPT–Total for 2 hypothetical individuals. The predicted value is represented by the horizontal line within the box; the prediction interval is represented by the box around the line. The increased precision with the use of regression and distribution-free intervals is seen both with the "usual" case in Panel A and the less typical case in Panel B. The differences in predicted values and the decreased variability for individuals with good initial scores are seen in Panel B.

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

residuals of the Halstead Index were skewed with a long right tail for individuals with a low Time 1 score. For these cases, probabilities based on the *z* score for the Halstead Index may be inaccurate.

DISCUSSION

This study addressed three main questions: which factors or variables should be considered in predicting follow-up results on neuropsychological tests, which statistical method is best for predicting the follow-up test scores, and which method is best for determining the likelihood that any given deviation from a predicted follow-up score represents a true change in ability? These questions were considered using seven illustrative test measures from the WAIS and the Halstead–Reitan Battery.

Some advantages of this study's design include (1) a participant sample that is significantly larger than those in previous studies of test–retest changes in neuropsychological performance, (2) inclusion of subjects with widely varying demographic characteristics and levels of baseline test performance (Tables 1 and 2), (3) test–retest intervals that also are variable and are fairly representative of intervals involved in clinical and research situations, and (4) inclusion of test measures that have been used extensively in neuropsychological clinical and research applications. A limitation is that WAIS and not WAIS–R IQs were included. Although these two versions of the Wechsler Intelligence Scales have similar psychometric properties and appear to perform similarly in retest situations (Matarazzo et al., 1980; Wechsler, 1981) they are not identical (Reitan & Wolfson, 1990) and the version used here admittedly is outdated. Indeed, with the publication of the WAIS–III in 1997, the WAIS–R may soon be considered outdated! Nevertheless, we would suggest that the exact version of the Wechsler included here is of limited importance, since the goals of our study were not test specific.

Our results indicate that the most powerful predictor of follow-up test performance is initial test performance. Initial scores alone accounted for 67% to 88% of the variance in follow-up test scores. By contrast, addition of other predictors in the multiple regression model resulted in increase in explained follow-up test score variance of from 0.8% to 8.5% (Table 3).

After considering the linear component of the relationship between baseline and follow-up test scores, the multiple regressions also included small but statistically significant nonlinear components for four of the five Halstead-Reitan Battery measures (but for neither of the WAIS IQs). It appears that the nonlinear influence is to predict less deviant follow-up scores when Time 1 scores are extreme. This effect is similar to, but goes farther than, the linear correction for regression to the mean. It is seen for exceptionally poor or good Time 1 scores on the AIR and Halstead Index, exceptionally good scores on Trails B and exceptionally poor scores on the TPT.

Demographic variables contributed significantly to the prediction of all follow-up test scores, even after baseline scores on the same tests were considered. In general, demographic variables tended to exert additional influences on follow-up scores that are in the same direction as their influences on initial scores (Heaton et al., 1996); for example, even given the same initial score on a test, older and less well educated participants tend to do somewhat worse on follow-up testing than do younger and better educated individuals. These findings with regard to the influence of age on test–retest changes are consistent with previous reports of reduced practice effects in older groups (Horton, 1992; Mitrushina & Satz, 1991; Ryan et al., 1992; Shatz, 1981). Similarly, in the current study, participants with worse overall neuropsychological competence at baseline (represented here by the Average Impairment Rating) tended to do even worse on second administrations of individual tests than would be predicted on the basis of their baseline scores on the same tests.

Test–retest interval had only a small (though still significant) influence for just three of the seven follow-up test measures. The relative lack of significance of time interval in the present study is consistent with the findings of McSweeny et al. (McSweeny et al., 1993), who reported minimal effects on the WAIS–R and Wechsler Memory Scale–Revised retest scores for 50 clinically stable patients with epilepsy. At least within the limited range of test–retest intervals considered here (2–16 months), it appears that practice effects do not decrease very much over time.

In sum, although initial performances on these tests were the best predictors of follow-up performances, other factors in multivariate models did increase prediction accuracy to some extent. The most important of these additional predictors were overall neuropsychological competence at baseline and demographic characteristics that are known to predict performances on these tests the first time they are administered. Unfortunately, indicators of the participant's state at the time of testing were not recorded consistently across the studies and, hence, were not included as potential predictors.

Of the four statistical approaches to predicting follow-up test scores, the simple Reliable Change Index (RCI) method clearly performed least well. This method is considered inadequate because of both its wide prediction intervals and its poor prediction accuracy (Table 5). Correction for practice effect (PE) does not affect the width of the prediction interval but helps considerably with the prediction accuracy. Indeed, in terms of overall prediction accuracy, results for the RCI + PE method are not much different from those of even the most complex (multiple regression) method. Figure 1 demonstrates, however, that large differences in predicted retest scores do occur for Models 2 to 4, especially at extremes of initial test performance and/or extremes of general neuropsychological competence at baseline.

The residual differences between predicted and obtained follow-up scores were essentially normally distributed for five of seven test measures. In these cases computation of

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

prediction intervals based upon the normal curve is both convenient and justified. For the measures that did have nonnormal residuals (TPT and Trails–B), however, the use of distribution free prediction intervals was quite important. In these cases, the width of the prediction intervals was greatly reduced, potentially allowing the detection of many more participants whose neurobehavioral functioning has changed over the follow-up period.

On all five of the Halstead-Reitan Battery measures, participants with poor initial performance (and/or other characteristics associated with poor performance) demonstrated greater variability in the differences between observed and expected scores at follow-up; Table 7. What this means is that, for best overall accuracy, prediction intervals need to be longer for people who do poorly at baseline (yielding better specificity), but can be reduced somewhat for the other participants (for better sensitivity). Please note that although level of initial performance is a convenient way to delineate the high variability subgroup, these cases have common characteristics that suggest these were not just random individuals who "had a bad day" at initial testing. The participants with poor initial performance tended to be older and less well educated than those with better initial performance.

Although the intervals are wider, in some instances than one might like, they actually represent a best case. Participants had the initial and following testing in the same setting, often with the same examiner. Also, since the testing was for research rather than for clinical purposes, there was little reason for subjects to be anxious about what the tests would show.

Interestingly, unlike the Halstead–Reitan Battery measures, WAIS IQs showed fairly constant variability in obtained-minus-predicted residuals at all levels of initial performance and demographics (Table 7). The likely reason for this is that, unlike the Halstead-Reitan Battery scores, the IQs have been age corrected and have undergone other transformations to improve (normalize) their distributional properties. In this regard, it is worth noting that most test measures used in neuropsychological assessments have *not* been subjected to demographic or distributional corrections. Thus, interpretation of repeated administrations of such tests frequently may benefit from use of variable prediction intervals (e.g., for older *vs.* younger individuals).

Again, simpler prediction models appear to work best with participants who are more typical in terms of baseline test performance and demographic characteristics. Simple models perform less well than do complex models with more impaired persons, and with those whose demographic characteristics tend to be associated with poorer absolute levels of performance. Clinical populations are likely to include many such individuals, so use of complex models may be indicated for them. It should be noted, however, that it is unwise to use any prediction model outside the range of values used to derive it. The ranges of demographic variables represented here are given in Table 1 while the ranges of initial scores are given in Table 2.

A decision to use more complex prediction models must weigh the models' potential advantages against the negative factors of increased workload and the potential for increased computation error rates. This is especially true if the computations will be done by hand (or calculator). Here is an area of practice in which computer software would be quite helpful, once the prediction models have been validated with clinical populations.

Probably the most important limitation of the current study is that it only included presumably healthy participants. It is uncertain how well our results will generalize to groups of people with neurologic disorders and other conditions (e.g., psychiatric and pain disorders) that may affect level, reliability, and stability of performance in test–retest situations (Bornstein et al., 1987; McCaffrey & Westervelt, 1995). Research is needed to assess the various prediction models and associated norms with multiple clinical populations, including those that have stable as well as progressive or resolving impairments. Some work to that end is currently in progress. To the extent that disorders are associated with relatively extreme test scores, the suggested advantages of more complex prediction models may be even greater in clinical than in nonclinical populations. It remains to be seen, however, whether norms for change, even if based upon complex models that take a variety of predictor variables into account, can generalize adequately from one population to another.

ACKNOWLEDGMENTS

This study was supported by Grants (HS04146) and (HS05304) both from the Agency for Health Care Policy and Research, and Grant (HD33677) from the National Institutes of Health–National Center for Medical Rehabilitation Research, all to S.S. Dikmen.

REFERENCES

Bornstein, R.A., Baker, G.B., & Douglass, A.B. (1987). Short-term retest reliability of the Halstead-Reitan Battery in a normal sample. *Journal of Nervous and Mental Disease*, *175*, 229–232.

Chelune, G., 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.

Heaton, R.K., Grant, I., & Matthews, C.G. (1991). *Comprehensive norms for an expanded Halstead-Reitan battery: Demographic corrections, research findings, and clinical applications.* Odessa, FL: Psychological Assessment Resources, Inc.

Heaton, R.K., Ryan, L., Grant, I., & 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.

Horton, A.M. (1992). Neuropsychological practice effects x age: A brief note. *Perceptual and Motor Skills*, *75*, 257–258.

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.

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press

Kleinbaum, D.G. & Kupper, L.L. (1978). *Applied regression analysis and other multivariate methods*. North Scituate, MA: Duxbury Press.

Matarazzo, J.D., Carmody, T.P., & Jacobs, L.D. (1980). Test–retest reliability and stability of the WAIS: A literature review with implications for clinical practice. *Journal of Clinical Neuropsychology*, *2*, 89–105.

McCaffrey, R.J. & Westervelt, H.J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*, 203–221.

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. *Clinical Neuropsychologist*, *7*, 300–312.

Mitrushina, M. & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology*, *47*, 790–801.

Reitan, R.M. & Wolfson, D. (1990). A consideration of the comparability of the WAIS and WAIS-R. *Clinical Neuropsychologist*, *4*, 80–85.

Reitan, R.M. & Wolfson, D. (1993). *The Halstead–Reitan Neuropsychological Test Battery: Theory and clinical interpretation*. (2nd ed.). Tucson, AZ: Neuropsychology Press.

Russell, E.W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach*. New York: Wiley.

Ryan, J.J., Paolo, A.M., & Brungardt, T.M. (1992). WAIS–R test–retest stability in normal persons 75 years and older. *Clinical Neuropsychologist*, *6*, 3–8.

Sawrie, S.M., Chelune, G.J., Naugle, R.I., & Lüders, H.O. (1996). Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. *Journal of the International Neuropsychological Society*, *2*, 556–564.

Shatz, M.W. (1981). WAIS practice effects in clinical neuropsychology. *Journal of Clinical Neuropsychology*, *3*, 171–179.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York: The Psychological Corporation.

Wechsler, D. (1981). *The Wechsler Adult Intelligence Scale–Revised manual*. New York: Harcourt, Brace, Jovanovich.

https://doi.org/10.1017/S1355617799544068 Published online by Cambridge University Press