Pergamon

Archives of Clinical Neuropsychology
16 (2001) 75–91

Archives
of
CLINICAL
NEUROPSYCHOLOGY

Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples

Robert K. Heaton[a,]*, Nancy Temkin[b], Sureyya Dikmen[b], Nanci Avitable[c], Michael J. Taylor[a], Thomas D. Marcotte[a], Igor Grant[a,d]

[a]*University of California at San Diego, San Diego, CA, USA*
[b]*University of Washington, Seattle, WA, USA*
[c]*University of Denver, Denver, CO, USA*
[d]*University of California at San Diego and V.A. San Diego Healthcare System, San Diego, CA, USA*

Accepted 22 November 1999

Abstract

Detecting change in individual patients is an important goal of neuropsychological testing. However, limited information is available about test-retest changes, and well-validated prediction methods are lacking. Using a large nonclinical subject group ($N=384$), we recently investigated test-retest reliabilities and practice effects on the Wechsler Adult Intelligence Scale and Halstead-Reitan Battery. Data from this group also were used to develop models for predicting follow-up test scores and establish confidence intervals around them. In this article we review those findings, examine their generalizability to new nonclinical and clinical groups, and explore the sensitivity of the prediction models to real change. Despite similarities across samples in reliability coefficients and practice effects, limits to the generalizability of prediction methods were found. Also, when multiple test measures were considered together, one or more "significant" changes were common in all (including stable) subject groups. By employing normative cut-offs that correct for this, sensitivity of the models to neurological recovery and deterioration was modest to good. More complex regression models were not more accurate than the simpler Reliable Change Index with correction for practice effects when confidence intervals for all methods were adjusted for variations in level of baseline test performance. © 2000 National Academy of Neuropsychology. Published by Elsevier Science Ltd.

* Corresponding author. Department of Psychiatry, University of California at San Diego, 3427 Fourth Avenue, San Diego, CA 92103.

PII: S0887-6177(99)00062-1

One of the major uses of neuropsychological (NP) testing is to measure change over time, including deterioration with progressive illnesses or improvements due to recovery from brain injury, or due to various kinds of treatments (Casey, Ferguson, Kimura, & Hachinski, 1989). The test-retest intervals may be short (e.g., in medication trials) or a year or more when one is looking for more gradual changes.

It is well-known that when NP tests are readministered, one cannot expect to get the same results even when there has been no real change in the patient. This is because tests are not perfectly reliable, and people naturally fluctuate in their functioning from time to time. Also, some tests and some people yield more variable results than others (Bornstein, Baker, & Douglass, 1987; Dodrill & Troupin, 1975; Feinstein, Brown, & Ron, 1994; Horton, 1992; Mitrushina & Satz, 1991). How does the clinician or researcher correct for these "noise" factors when trying to measure real changes in neurologic status?

In research comparing group mean changes, statistical methods are fairly straightforward if one has an appropriate control group. However, this does not help the clinician who needs to determine, for an individual case, whether given fluctuations on various tests represent meaningful changes or normal variability in performance. Even in research, group mean comparisons may not tell the whole story, because a treatment or clinical condition may cause scattered changes in the functioning of only a subset of people who are being evaluated. Here too, it would be important to be able to confidently detect and quantify real changes on individual patients or subjects.

As important as these decisions are, our technology for making them is not well-worked out and the needed norms are even more lacking. In just the last few years, research groups in the epilepsy area have begun to address this problem, using two general approaches to the prediction of follow-up scores in neurologically stable people (Chelune, Naugle, Luders, Sedlak, & Awad, 1993; Hermann et al., 1991; McSweeny, Naugle, Chelune, & Luders, 1993; Sawrie, Chelune, & Naugle, 1996). The first is an adaptation of the Reliable Change Index (RCI) from psychotherapy research (Jacobson & Truax, 1991): The predicted follow-up score is simply the baseline score, possibly adjusted by a constant that is the mean practice effect from some reference group. The other approach to predicting follow-up scores uses a regression model developed from test-retest results of a stable reference group (McSweeny et al., 1993). Then, using the same normative data, confidence intervals (CIs) are established around the predicted score. With the typical 90% CI, only 5% of neurologically stable people would be expected to score outside of the interval in each direction. If a new patient scores on retest within the CI, he/she is considered to be unchanged. If the person scores outside of the interval, he/she is considered to have shown a significant improvement or deterioration, depending on the direction of the change. This approach should be valid, as long as the prediction equation and CI are based on a reference group that is appropriate to the new person being examined.

The research programs cited above have been directed at detecting meaningful changes in NP functioning after temporal lobectomies for intractable epilepsy. The reference groups in each case were 40 to 50 nonsurgical patients with intractable epilepsy, having test-retest intervals averaging from 6 to 9 months. Appropriate norms were provided for two test batteries that are commonly used in epilepsy surgery programs. These have been major advances but, as the authors pointed out, many questions remain about the determinants and sizes of test-retest changes in different people, the best statistical methods for modeling such changes, and how well norms for change generalize from one sample or population to another.

Recently, we used retest results of a large sample of neurologically normal or stable adults ($N=384$) to explore the nature and predictors of change on selected measures from the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) and Halstead-Reitan NP Battery (HRB; Dikmen, Heaton, Grant, & Temkin, 1999). Although a minority of the sample had a history of prior alcohol-use disorder (15%) and/or a prior traumatic head injury (7%), these were not recent problems that would be expected to cause real changes in subjects' functioning over the test-retest intervals. Retest reliability estimates for the test measures were generally good (.70 to the low .90s). Practice effects were variable across tests and tended to be larger on measures with significant problem-solving and novelty components. Also, the amount of test-retest change was not uniform across subjects. The findings suggest that practice effects may vary somewhat with test-retest interval as well as with subject characteristics. Regarding the latter, younger subjects, those with more education and those with higher general NP competence at baseline were likely to show greater practice effects. Finally, initial high and low scorers, respectively, showed less versus more improvement, consistent with regression to the mean.

In a second article, we explored the use of different prediction models and CIs for detecting "real" change in the same subject group (Temkin, Heaton, Grant, & Dikmen, 1999). The specific prediction models were the simple RCI method, the RCI with correction for practice effects, a simple regression model that predicts follow-up from baseline scores, and a multiple regression model that includes other potential predictors (retest interval, demographic factors, prior neurological history, general NP competence at baseline, and nonlinear effects). Initial test score was by far the most powerful predictor of all follow-up scores, although prediction was sometimes improved slightly by considering other variables (primarily demographics and general NP competence). The simple RCI method performed least well because of its failure to correct for practice. The other three models had comparable and higher overall prediction accuracies, but provided different predictions at extremes of initial performance and demographic characteristics. Prediction error (and resulting size of CIs) also tended to be higher among subjects with relatively poor initial performances. The latter finding may be particularly important when attempting to detect real change in clinical groups because they are more likely to include individuals with impaired initial performance. This also raises concerns about whether norms for change, no matter what statistical model they are based on, can be generalized from nonclinical to clinical groups.

In view of the above findings, Temkin et al. (1999) tentatively recommended use of the more complex multiple regression models, with attention to differential variability (and associated required CIs) in good versus poor initial performers. It was suggested that simpler models will perform equally well for people who are more "average" in terms of initial performance and demographic characteristics, but that more complex models would be expected to do better with extreme cases that are apt to be more common in clinical samples. Nevertheless, it was acknowledged that research was needed to determine the generalizability of the proposed norms for change, to both nonclinical and clinical populations.

To assess the generalizability of the methods and norms presented by Temkin et al. (1999) and to determine their sensitivity to change in neurologic status, we have applied them to four new subject groups: a nonclinical cross-validation group, a stable clinical group of persons with chronic schizophrenia, a group of people recovering from recent moderate-to-severe traumatic brain injuries (TBIs), and a small group of people who suffered new brain insults or other sources of brain compromise between baseline and follow-up testing. Since the simple RCI prediction model was found to provide substantially less accurate predictions with the original Temkin et al. (1999) sample, the current presentation compares results of only the other three models: RCI + practice, simple regression, and multivariate regression. Variable confidence intervals for high- versus low-baseline performers are applied with all three prediction models. The test variables are the same as in the Temkin et al. (1999) study: WAIS Verbal and Performance IQs and five measures from the HRB (Category Errors [Category], Trails B–Time [Trails B], Tactual Performance Test [TPT] time per block, the Halstead Impairment Index [HII], and the Average Impairment Rating [AIR]).

The following questions are addressed: How well do the Temkin et al. (1999) sample (hereafter called "Base" sample) estimates of reliability and practice effects generalize to other neurologically stable groups? How well do Base-sample predictions (regarding follow-up scores) generalize to new stable groups? Which individual test measures are best for measuring change, and how good are they? In attempting to detect neurological change, does it help to consider multiple test measures together? How sensitive and specific are the three prediction models in detecting change?

1. Method

1.1. Subjects

The cross-validation study included one new normal group and one "stable" clinical group. The former consisted of 124 healthy, uninfected subjects who were serving as controls for a longitudinal investigation of the neurobehavioral sequelae of HIV-1 infection. The clinical group was composed of 69 subjects with schizophrenia who were being followed in longitudinal investigations of that disorder at the University of California at San Diego. Subjects in both groups had no history of neurologic disease or significant brain injury and no alcohol-use disorder within 6 months of either testing. The schizophrenia subjects were clinically stable outpatients, whose diagnoses were established based on the Structured Clinical Interview for the *DSM-III-R* (SCID; Spitzer & Williams, 1986). Table 1

Table 1
Demographic characteristics and test-retest interval for three stable groups

	Group		
	Base normal ($n=358$)	Cross-validation normal ($n=124$)	Schizophrenic ($n=69$)
Mean age	34.7	40.1	40.5
Mean edcucation	12.3	14.3	13.1
% Male	65.6	64.5	65.2
% Caucasian	88.3	49.2	85.5
Mean interval	9.1	13.6	16.0

describes the Base sample of subjects from Temkin et al. (1999) who had complete data on all test measures, along with the two new cross-validation groups. The Base normal subjects were somewhat younger, had less education, and had a somewhat briefer test-retest interval than subjects in the other new groups. The cross-validation normal group included more ethnic minority subjects. (If such differences in demographics and retest intervals substantially affect the generalizability of the Base sample results, their value as norms obviously would be limited.)

The two "change" groups consisted of (a) 23 subjects who underwent baseline testing 1 month following a moderate-to-severe TBI with coma lasting at least 3.5 days (Recovering TBI), and (b) 10 subjects who had histories of new brain compromise between baseline and follow-up testing (New Insult). The subjects in the acute trauma group were participating at the University of Washington in a longitudinal study of recovery from TBI. The New Insult group consisted of subjects who were either clinical patients ($n=2$) or participants in longitudinal studies at the University of California at San Diego concerning alcoholism ($n=6$) or HIV-1 infection ($n=2$); between the baseline and follow-up testings, 5 of the subjects suffered cerebral vascular accidents, 1 suffered a traumatic brain injury requiring surgical evacuation of a subdural hematoma, 2 suffered independently documented neurological complications of HIV-1 infection (encephalitis, dementia), and 2 evidenced clear clinical progression of Alzheimer's disease. Table 2 describes the two "change" groups. Both differed from the Base normal group with respect to age (one group was younger, the other older). Also, the New Insult group included more males and had a longer test-retest interval.

1.2. Test measures

The NP variables used both by Temkin et al. (1999) and here were selected for demonstration purposes because they are frequently used, representative measures from the WAIS and HRB. As noted above, they included the WAIS Verbal and Performance IQs, two summary scores from the HRB (HII and AIR) and three individual test scores from that battery (Category, Trails B, and TPT time per block).

All subjects had complete baseline and follow-up data on all test measures, and in no case was a different version of a test administered at follow-up than at baseline. However, the Base normal group and the Recovering TBI group were administered the WAIS, whereas the two cross-validation stable groups received the Wechsler Adult Intelligence Scale-Revised

Table 2
Demographic characteristics and test-retest interval for two change groups

	Group	
	Recovering TBI ($n=23$)	New insult ($n=10$)
Mean age	24.6	46.7
Mean education	12.9	13.5
% Male	69.6	90.0
% Caucasian	95.7	90.0
Mean interval	11.0	35.0

Note. TBI = traumatic brain injury.

(WAIS-R; Wechsler, 1981); half of the subjects in the New Insult group received each version of the WAIS. Clearly, this is a potential limitation of the study. However, although mean IQ levels tend to be somewhat lower for the WAIS-R than for the WAIS, our focus here is on change over time; also, the test-retest reliabilities and mean practice effects are quite comparable for these two versions of the WAIS. This similarity is borne out in comparably low standard errors of measurement for the WAIS (Base sample) and WAIS-R (cross-validation normal group). In the current study, respectively, these standard errors are 3.4 and 3.6 for the Verbal IQ, and 4.3 and 4.7 for the Performance IQ. Furthermore, Fig. 1 demonstrates comparable test-retest changes (parallel lines) for the respective normal groups on the WAIS and WAIS-R. Finally, as will be seen, IQ predictions from the Base-sample data generalize as well across groups, as do summary measures from the HRB (which do not have different versions).

1.3. Procedures

Details of the prediction models and CI determinations can be found in Temkin et al. (1999). Briefly, for the RCI + practice model, the predicted follow-up test score for any subject is that person's baseline score plus the mean practice effect of the Base sample on that test measure. The 90% CI for that estimate is the predicted score $\pm(1.64 \times SE_{\text{diff}})$, where SE_{diff} is the standard deviation of the test-retest differences of the Base sample, or corresponding subgroups of the Base sample (see below).

In the simple regression model, the predicted follow-up score of an individual comes from the equation generated when follow-up scores were regressed on baseline scores of the Base sample. In the multiple ("full") regression model, the baseline score always entered, and typically accounted for most of the variance in follow-up scores. However, three to six additional predictors were included in various combinations for the seven test variables (see

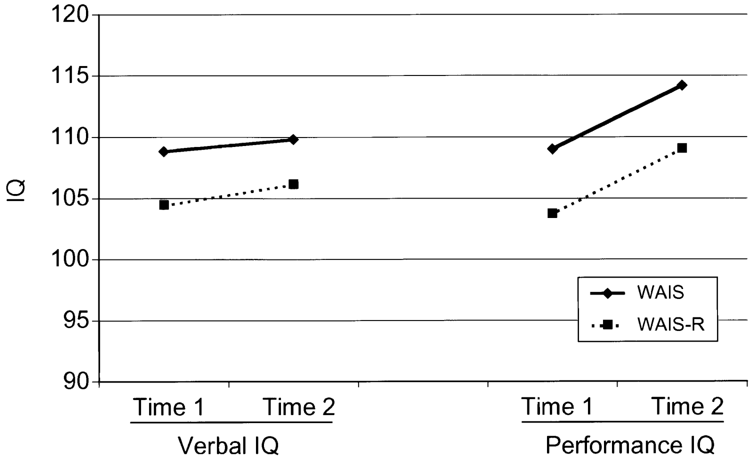

Fig. 1. Mean IQ changes between baseline (T1) and follow-up (T2) for the Base sample on the WAIS (Wechsler Adult Intelligence Scale) and for the cross-validation sample on the WAIS-R (Wechsler Adult Intelligence Scale-Revised).

Temkin et al., 1999, p. 362, Table 4). For both regression models, the 90% CI was defined by the predicted score $\pm(1.64 \times SE_{\mathrm{residual}})$, where the latter term is the standard deviation of the residual from the regression equation for either the entire Base sample or the corresponding subgroup of the Base sample (see below).

An exception to the above generalizations about determination of CIs was necessary in some instances. As already noted, Temkin et al. (1999) found that prediction errors (and therefore optimal CIs) sometimes varied according to subjects' initial levels of test performance. That is, variability in test-retest differences sometimes were much higher for subjects with poor baseline performances. For three variables (Category, Trails B, and TPT), this effect was substantial across all prediction models, necessitating determination of different CIs for subjects with poor versus average (or better) baseline performance. After inspection of the data, we arrived at the following operational definitions of low baseline performances on the respective tests: ≥ 60 errors on Category, ≥ 92 seconds on Trails B, and ≥ 0.64 minutes per block on TPT. To illustrate this point, Table 3 presents the SE_{diff} values obtained for the three test variables with subgroups divided on the basis of initial levels of performance.

The procedures just discussed were used to determine the 90% CIs for each test measure within each model. For individual tests, the stable groups should contain approximately 90% of subjects within the "same" (no change) category and about 5% each within "better" and "worse" categories; any significant deviations from normal and schizophrenic groups would reflect less-than-desirable generalizability of the Base sample's normative findings. With the two change groups, sensitivity to real change would be indicated by higher percentages of TBI subjects being classified as better, and of New Insult subjects being classified as worse.

We also were interested in exploring whether improved clinical classifications (specificity and sensitivity) can be achieved by using results of multiple test measures together. For this purpose, we determined for each subject within the Base sample the numbers of follow-up test scores (out of the possible seven) classified as either better or worse by the CI criteria described above. We then computed better-worse difference scores (i.e., number of better scores minus number of worse scores) for each subject and examined the distribution of these difference scores within the Base sample. For all three prediction models, approximately 90% of subjects obtained difference scores between $+1$ and -1; this range therefore defined overall performance across the seven measures that was unchanged (same). Approximately equal numbers of Base subjects obtained difference scores $\geq +2$ and ≤ -2, and these respective ranges defined better or worse performance across the seven measures. These

Table 3
Standard errors of the difference between time 1 and time 2 scores for subgroups with $\geq$ average versus low baseline scores

	Test variables		
Baseline score	Category SE_{diff}	Trails B SE_{diff}	TPT SE_{diff}
$\geq$ Average	11.0	14.9	0.09
Low	20.1	41.5	0.42

Note. TPT = Tactual Performance Test.

operational definitions of overall stability or change on the test battery were used to compare the three prediction models with respect to diagnostic accuracy with subjects in the two new stable groups and two change groups.

Finally, for some of the analyses and presentations below, raw test scores were converted to age-, education-, and gender-adjusted T scores, using procedures described in Heaton, Grant, and Matthews (1991) and Heaton (1992). These conversions place all scores on a common metric and enable more direct comparisons of groups that differ on those demographic characteristics.

2. Results and discussion

Table 4 summarizes the baseline and follow-up test results of all five subject groups. It was expected that the two normal groups would outperform the clinical groups at both evaluations, and that performances of all except the New Insult group would show improvement from baseline to follow-up. The latter improvement would represent practice effect for the three stable groups, and perhaps some combination of practice effect and recovery in the TBI group. The New Insult group was expected to show deterioration of functioning on the follow-up testing.

These predictions were tested using two repeated measures analyses of variance (ANOVAs) of mean T scores (averaged across the seven NP measures). The first of these

Table 4
Baseline and follow-up mean and standard deviation scores for three stable and two change groups

	Stable groups						Change groups			
	Base normal ($n = 358$)		Cross-validation normal ($n = 124$)		Schizophrenic ($n = 69$)		Recovering TBI ($n = 23$)		New insult ($n = 10$)	
Test measure	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
Baseline										
Verbal IQ	108.8	(13.8)	104.4	(11.9)	96.7	(12.7)	93.3	(13.9)	102.2	(7.3)
Performance IQ	109.0	(11.4)	103.7	(12.2)	95.4	(13.4)	90.5	(11.3)	105.5	(10.4)
HII	0.3	(0.3)	0.3	(0.2)	0.6	(0.3)	0.6	(0.3)	0.5	(0.3)
AIR	1.0	(0.5)	1.0	(0.5)	1.7	(0.7)	1.7	(0.6)	1.5	(0.6)
Category	40.1	(25.7)	39.8	(24.6)	65.2	(34.4)	47.6	(20.1)	43.7	(25.3)
Trails B	70.6	(42.5)	65.3	(24.8)	107.9	(61.7)	117.7	(53.3)	102.2	(64.2)
TPT	0.5	(0.4)	0.5	(0.3)	1.3	(1.6)	1.0	(1.1)	0.6	(0.3)
Follow-up										
Verbal IQ	109.8	(14.0)	106.1	(12.5)	98.1	(13.2)	101.2	(12.6)	98.1	(11.8)
Performance IQ	114.2	(12.5)	109.0	(13.9)	97.7	(14.8)	107.3	(10.1)	97.0	(13.1)
HII	0.2	(0.3)	0.2	(0.2)	0.5	(0.3)	0.3	(0.2)	0.7	(0.3)
AIR	0.9	(0.5)	0.9	(0.5)	1.5	(0.7)	1.0	(0.4)	2.2	(1.0)
Category	29.5	(24.4)	29.6	(21.0)	48.2	(32.0)	29.1	(17.8)	58.7	(29.6)
Trails B	66.6	(42.7)	66.8	(26.5)	97.4	(62.5)	62.6	(22.3)	152.7	(88.6)
TPT	0.4	(0.3)	0.5	(0.3)	1.1	(1.6)	0.4	(0.2)	1.8	(2.9)

Note: TBI = traumatic brain injury; HII = Halstead Impairment Index; AIR = Average Impairment Rating; TPT = Tactual Performance Test.

compared the three stable groups. As expected, a significant group effect was obtained, $F(2, 548) = 43.90$, $p < .001$, in addition to a significant time effect, $F(1, 548) = 301.93$, $p < .001$; follow-up contrasts revealed that the two normal groups outperformed the schizophrenic group, and that follow-up scores were better than baseline scores. Unexpectedly, a small but significant group × time interaction effect was also obtained, $F(2, 548) = 7.49$, $p < .001$, due to the fact that the cross-validation normal group evidenced slightly less practice effect than either the Base normal group or the schizophrenic group (see Fig. 2).

It is unclear why the cross-validation normal group showed a slightly smaller practice effect than the other two stable groups. Although its mean test-retest interval is longer than that of the Base normal group, it is shorter than that of the schizophrenic group (see Table 1). Another possibility is that the much higher percent of ethnic minorities (mostly African Americans) in the cross-validation normal group might account for the group differences in practice effects. This, however, does not appear to be the case because the White and non-White subjects within the cross-validation normal group had comparable mean practice effects: mean T scores of 2.94 versus 2.10, respectively, $t(122) = 1.30$, $p = .20$. Of perhaps more interest is the finding that despite the schizophrenic group's NP impairment at both baseline and follow-up testings, it evidenced an overall mean practice effect that was substantial and almost identical to that of the Base normal group (see Fig. 2). Thus, within a fairly broad range of initial test performance, baseline levels do not appear to greatly influence practice effects on retest.

Our second repeated measures ANOVA compared the Base normal group to the two change groups. Again, as expected, we obtained significant effects for group, $F(2, 388) = 23.10$, $p < .001$, and time, $F(1, 388) = 38.04$, $p < .001$. Also, as was expected in this case, the group × time interaction effect was quite robust, $F(2, 388) = 100.68$, $p < .001$. Fig. 2

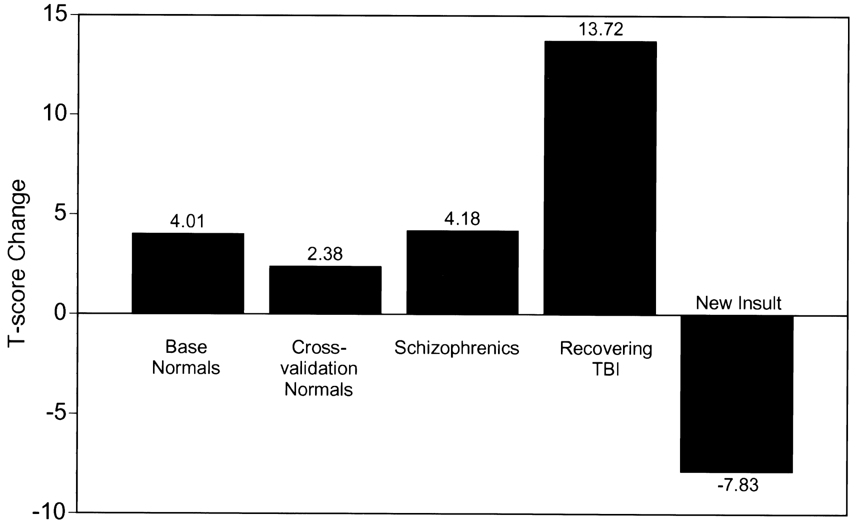

Fig. 2. Mean T-score changes (follow-up minus baseline) for five subject groups, averaged across test measures. TBI = traumatic brain injury.

Table 5
Test-retest reliability coefficients and standard errors of the difference for three stable groups

	Group					
	Base normal ($n=358$)		Cross-validation normal ($n=124$)		Schizophrenic ($n=69$)	
Test measure	r	SE_{diff}	r	SE_{diff}	r	SE_{diff}
Verbal IQ	.94	4.8	.91	5.0	.92	5.1
Performance IQ	.86	6.0	.85	6.7	.88	6.6
HII	.80	0.2	.70	0.2	.80	0.2
AIR	.92	0.2	.88	0.3	.91	0.3
Category	.84	14.5	.77	16.7	.77	23.4
Trails B	.87	21.7	.72	18.6	.77	41.9
TPT	.87	0.2	.67	0.2	.71	1.2

Note. HII = Halstead Impairment Index; AIR = Average Impairment Rating; TPT = Tactual Performance Test.

shows the modest practice effects for the three stable groups, the substantial additional recovery of function in the TBI group, and the significant worsening of the New Insult group. Note that the actual "loss" suffered by the New Insult group is the sum of the negative component in the figure plus the expected practice effect (i.e., the practice effect that would be expected in the absence of a new insult).

2.1. Detecting absence of neurologic change: specificity

In this section, we will consider how well results of our NP measures and Base-sample prediction models generalize to the two new groups of stable subjects: the cross-validation normal group and the schizophrenic group. First, however, it was of interest to determine whether the two new stable groups showed comparable test-retest reliability and comparable variability of NP test-score changes over time. These data are reported in Table 5. The four NP summary scores showed comparably high test-retest correlations and low variability of the Time 1 minus Time 2 difference scores. However, although the schizophrenic group's test-retest reliability estimates on the individual test measures remained acceptably high, its standard errors on those measures were considerably worse than those for the two normal groups. These standard errors are equal to or larger than those of the low-functioning subgroup of the Base sample as shown in Table 3. This finding of large standard errors with the schizophrenic group suggests that CIs developed with normal subjects may not generalize well to impaired populations, even if they have remained neurologically stable over the test-retest intervals.

Table 6 reports the accuracy with which Base sample prediction models, on average, detected "no change" in the two cross-validation groups. Because we are using 90% CIs, a perfect cross-validation would be to classify 90% of subjects in these groups as unchanged. The table shows that the Base sample norms for change performed quite well with the cross-validation normal group on all test measures. However, substantial increases in error rates are seen with the schizophrenic group subjects, particularly for the Performance IQ and the individual HRB tests.

Table 6
Percentages of subjects in stable crossvalidation groups correctly classified as unchanged by individual test measures, averaged across prediction models

	Group	
Test measure	Cross-validation normal (*n* = 124)	Schizophrenic (*n* = 69)
Verbal IQ	88.2	89.4
Performance IQ	86.8	79.7
HII	89.5	84.6
AIR	86.5	82.1
Category	88.7	78.8
Trails B	85.8	79.2
TPT	87.9	78.3

Note: HII = Halstead Impairment Index; AIR = Average Impairment Rating; TPT = Tactual Performance Test.

Table 7 displays the accuracies of the three prediction models in classifying subjects within these groups, using the seven NP test measures together. As detailed in the Procedures section above, subjects were classified as unchanged on the total battery if the differences between their "significantly" positive and negative test-score changes (out of 7) were between +1 and −1. The table indicates that for all three prediction models, the Base sample norms performed substantially better for the cross-validation normal group than for the schizophrenic group (with respective mean "error" rates of 5.6% vs. 21.9%). Furthermore, within the cross-validation groups, there are no substantial differences in the accuracies of the three prediction models.

The above results clearly indicate increased error rates when attempting to use Base sample norms to correctly classify schizophrenic patients as neurologically unchanged. This was perhaps foreshadowed by the earlier finding of increased test-retest variability in schizophrenic, relative to normal subjects, on some NP measures. A possible cause of such increased variability might be fluctuations in the schizophrenic patients' psychiatric symptoms. To explore this possibility, we examined the clinical symptom changes of subgroups of schizophrenic subjects who were classified as NP-improved ($n=11$), NP-same ($n=48$), and NP-worse ($n=10$) by the RCI + practice model. (Again, the other prediction models yielded similar results.) The three NP-defined subgroups were compared with respect to their test-retest changes on two widely used measures of psychotic symptoms: the Scale for Assessment of Negative Symptoms (SANS; Andreasen, 1983) and the Scale for Assessment

Table 7
Percentages of subjects in stable crossvalidation groups correctly classified as unchanged by three prediction models, using decision rules for combined test battery

	Group	
Prediction model	Cross-validation normal (*n* = 125)	Schizophrenics (*n* = 69)
RCI + practice	83.9	69.6
Simple regression	86.3	66.7
Full regression	83.1	68.1

Note: RCI = Reliable Change Index.

of Positive Symptoms (SAPS; Andreasen, 1984). These analyses excluded 5 of the 69 subjects because they were missing clinical symptoms ratings at the time of one or both NP testings. ANOVAs for both symptom rating scales yielded nonsignificant main effects for NP groups, SANS $F(2, 61) = 0.50$, SAPS $F(2, 61) = 0.39$, and for time, SANS $F(1, 61) = 0.01$, SAPS $F(1, 61) = 1.82$. Group × time interaction effects also were nonsignificant, SANS $F(2, 16) = 0.01$, SAPS $F(2, 61) = 0.81$. These results indicate relative stability of psychotic symptoms for all three NP-defined subgroups, and suggest that the greater-than-expected number of schizophrenic patients who appeared to show NP-worsening or NP-improvement (by standards developed with normal controls) were not due to fluctuations in the patients' clinical symptoms.

It appears that even in schizophrenic subjects who remain stable clinically, test-retest changes in NP functioning may be larger (in both better and worse directions) than is the case with normal subjects. Furthermore, this is true despite the fact that we used much larger CIs for subjects who initially performed poorly on the three tests that required variable CIs in normal subjects (Category, Trails B, and TPT; as noted above, on these measures, Base normal subjects with low initial scores had a substantially higher SE_{diff} value than those with average or better initial scores).

Research is needed to determine whether this relative failure to generalize from normal to schizophrenic subjects will be seen with other psychiatric and neurologic patient groups. Also, because the NP fluctuations in our schizophrenic subjects cannot be explained by changes in their psychotic symptoms, it is unclear what factors necessitate the use of wider CIs. For example, is it the fact that more of them had poor baseline performance, and that some of their impairments were greater than those of the worst performing normal subjects? (If so, norms for change may still generalize well across clinical groups that have comparable baseline impairments.) Unfortunately, our total sample size of only 69 schizophrenic patients greatly limits our ability to explore these issues with the data at hand; for example, only about seven more than the expected numbers of subjects were classified as better or worse, respectively.

If separate norms for change *are* needed for different clinical and nonclinical groups, progress in this area will likely be slower and more expensive. In addition, it should be kept in mind that any need for larger CIs around NP predictions in clinical groups will be likely to reduce the sensitivity of the tests to true change in neurobehavioral status (especially if the real change is subtle or mild).

2.2. Detecting neurologic change: sensitivity

Here we consider how well the tests and prediction models detect improvement or worsening of brain functions using data from the Recovering TBI and New Insult groups. A change classification will be considered "correct" whenever a patient in the TBI group is called "improved on retest" and when a New Insult patient is called "worse on retest." This generalization seems apt for the New Insult group because all of these individuals had documentation of new or increased central nervous system disorders from baseline to retest. However, whereas most people with moderate-to-severe TBIs can be expected to get better from 1 month to 1 year postinjury, clinically meaningful improvement may not occur in all

Table 8
Percentages of subjects in change groups correctly classified as improved (recovering TBI) or worse (new insult) on individual test measures, averaged across prediction models

	Group	
Test measure	Recovering TBI (*n* = 23)	New insult (*n* = 10)
Verbal IQ	40.6	26.7
Performance IQ	49.3	50.0
HII	36.2	36.7
AIR	59.4	80.0
Category	20.3	40.0
Trails B	40.6	60.0
TPT	34.8	56.7

Note: TBI = traumatic brain injury; HII = Halstead Impairment Index; AIR = Average Impairment Rating; TPT = Tactual Performance Test.

cases; thus, a perfectly sensitive system or test may not necessarily classify 100% of such TBI patients as improved on retest. Another caveat that should be noted before considering the results of these two change groups is that they are quite small. Thus, only tentative conclusions may be drawn from their data.

Table 8 reports the percentages of patients in the two groups who were identified by the individual test measures as changed in the expected direction on retest. These results are averaged across the three prediction models. None of the measures approached 100% accuracy with either group. In both groups, the AIR showed the greatest sensitivity to change, and this measure was especially accurate in detecting new or progressive central nervous system compromise. The HII, another HRB summary measure, did much less well at detecting change. This likely resulted from its restricted range of possible values and the dichotomous (impaired vs. unimpaired) nature of its contributing test scores. The Verbal IQ was less sensitive to change than the Performance IQ, a finding that is not surprising given the "crystallized" versus "fluid" nature of these respective measures of intelligence. Finally, the data suggest that most HRB measures are relatively more sensitive to worsening than to improving conditions. Again, however, it is possible that some TBI patients were classified as unchanged because they did not evidence clinically significant improvement in their neurobehavioral status from baseline to follow-up (see further analysis of this possibility below).

Table 9 presents the correct classification rates achieved by the three prediction models for the two change groups. As was the case for the above comparisons of the three models, the

Table 9
Percentages of subjects in change groups correctly classified as improved (recovering TBI) or worse (new insult) by three prediction models, using decision rules for combined test battery

	Group	
Predicton model	Recovering TBI (*n* = 23)	New insult (*n* = 10)
RCI + practice	73.9	80.0
Simple regression	73.9	90.0
Full regression	65.2	70.0

Note: TBI = traumatic brain injury; RCI = Reliable Change Index.

seven test variables were considered together in these analyses. That is, the number of unusual negative changes was subtracted from the number of unusual positive changes on the seven test measures, and any subject with a difference score between + 1 and − 1 was classified as unchanged; subjects with scores greater than + 1 were called improved, whereas those with scores less than − 1 were called worse.

As was the case with the specificity analyses (Table 7), there is not much difference among the three models in terms of overall correct classification rates. Also, most subjects in both groups were classified as changed in the expected directions, and the models performed slightly (although nonsignificantly) better with the New Insult group than with the Recovering TBI group. Not shown in Table 9 is that the classification "errors" for all models and both groups almost always were that they called subjects in the change groups "unchanged"; the single exception is that the full regression model classified one of the New Insult subjects as improved.

In detecting change, the models using the combined group of tests (Table 9) generally outperformed the individual test measures (Table 8). This was particularly true for the Recovering TBI group, in which the median correct classification of the individual test measures was only 40.6%. This reflects the fact that subjects in this group often showed "spotty" recovery on different groups of measures. On the other hand, as noted above, it is possible that not all of these TBI subjects actually improved from 1 month to 12 months postinjury. To explore this further, Fig. 3 depicts the profile of T-score changes (Time 2 minus Time 1) across the seven test measures for the Base normal group and two TBI subgroups: those classified as improved ($n = 17$) and unchanged ($n = 6$) by the RCI + practice model. The figure shows that the change profile of the TBI same subgroup closely resembles that of the

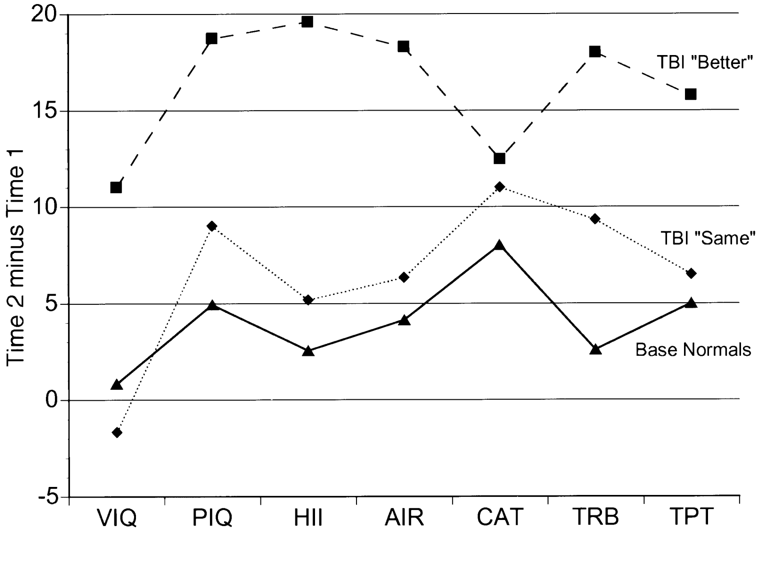

Fig. 3. Mean T-score improvements of the Base normal group compared to those for TBI cases classified as better ($n = 17$) and same ($n = 6$) by the RCI + practice model. TBI = traumatic brain injury; RCI = Reliable Change Index; VIQ = Verbal IQ; PIQ = Performance IQ; HII = Halstead Impairment Index; AIR = Average Impairment Rating; CAT = Category; TRB = Trails B; TPT = Tactual Performance Test.

Base normal group. Although the changes were in the positive direction, these were within the range of typical practice effects. Moreover, in this subgroup of TBI subjects, there was not a single instance of a significant improvement on any of the seven test measures. The reason for this is not entirely clear, although inspection of the baseline scores reveals that half of the subjects in this small subgroup already were within normal limits on the test battery 1 month after their injuries, and therefore may have essentially completed their recoveries at that early time. Another possibility is that the 6 TBI subjects who were classified as unchanged may have evidenced improvement on NP variables that were not considered here (e.g., measures of learning and memory).

The above observations suggest that our estimates of the NP models' sensitivity to improvement may be somewhat conservative. Especially if this is true, the current results support the tentative conclusion that the NP prediction models, when applied to multiple measures together, can be reasonably sensitive to both positive and negative changes in neurologic status. However, more secure conclusions await replication with larger samples of clinical populations of interest. Ideally this research would include subjects with a wider variety of progressive/new or resolving conditions, including those involved in pharmacologic and other treatment trials.

The WAIS/WAIS-R and HRB variables included in the current study are widely used and arguably representative NP measures. However, the goals of this study were not test specific. Rather, our purpose was to clarify general issues regarding the modeling of NP test-retest results and detecting presence or absence of meaningful changes in individual subjects. It is anticipated that future research of this kind will use different combinations of test measures that are tailored to the populations being investigated and the types of neurobehavioral changes being anticipated.

3. Conclusions

This study employed three statistical approaches for modeling NP test-retest results, and has compared the specificity and sensitivity of the models in detecting neurologic change. Although the predictions of the NP models generalized well from the Base normal group to a new group of normal subjects, larger-than-expected numbers of stable schizophrenic subjects were misclassified as NP-changed on retest (both in the positive and negative directions). This suggests that norms for change may not generalize adequately from nonclinical to clinical groups. Instead, better specificity generalization is likely to occur with stable normative samples that have initial levels of functioning comparable to those of individuals or populations with which the norms will be used.

Thus, very different kinds of subjects may be needed to provide the best norms for change versus norms for diagnostic purposes. The latter type of norms require neurologically normal subjects who are demographically similar to the new person or group to be evaluated, whereas norms for change may require neurologically stable (but not necessarily normal) subjects who have similar baseline test performance to that of the new person or group.

Although detection of NP change in clinical groups may require larger CIs to achieve specificity (correctly classifying stable subjects as unchanged), large CIs will necessarily make it more difficult to detect real change when it occurs. The challenge is apt to be especially great when the change to be detected is relatively mild and/or spotty (involving different patterns of abilities in different people).

In identifying real change within our New Insult group, and especially within our Recovering TBI group, the individual NP test measures did less well than the seven test scores together. On the other hand, use of multiple test measures increases the probability of finding one or more "unusual" changes in neurologically stable people. To adequately interpret such results, the neuropsychologist needs to know how many changes of a given magnitude are actually rare when the specific test battery is applied to a relevant population of neurologically stable subjects. Thus, norms for change are needed not only for individual test measures, but also for groups of tests that will be used together.

Despite theoretical advantages of the more complex regression-based prediction models over the RCI + practice model (Temkin et al., 1999), in the current study these advantages did not lead to better performance. When predictions based on the three models were applied to new groups of neurologically stable and changing subjects, the overall correct classification rates were not substantially different across models. Thus, for classification and interpretation of NP change scores, our results would support continued use of the simpler approach: RCI + practice, with CIs being adjusted as needed for different levels of baseline test performance.Dodrill Troupin 1975

Acknowledgments

This study was supported by Grants HS04146 and HS05304 from the Agency for Health Care Policy and Research, and Grant HD33677 from the National Institute of Health–National Center for Medical Rehabilitation Research; and by Grant P50 MH45294 from the National Institute of Mental Health. The authors thank Mariana Cherner, Lauren Dawson, Julie Gladsjo, and Barton Palmer for assistance in creating data sets and designing figures, and Carol Demong for help in preparing the manuscript.

References

Andreasen, N. C. (1983). *The scale for the assessment of negative symptoms (SANS)*. Iowa City, IA: University of Iowa.

Andreasen, N. D. (1984). *The scale for the assessment of positive symptoms (SAPS)*. Iowa City, IA: University of Iowa.

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

Casey, J. E., Ferguson, G. G., Kimura, D., & Hachinski, V. C. (1989). Neuropsychological improvement versus practice effect following unilateral carotid endarterectomy in patients without stroke. *Journal of Clinical and Experimental Neuropsychology, 11*, 461–470.

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

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Battery. *Journal of the International Neuropsychological Society, 5*, 346–356.

Dodrill, C. B., & Troupin, A. S. (1975). Effects of repeated administrations of a comprehensive neuropsychological battery among chronic epileptics. *The Journal of Nervous and Mental Disease, 161*, 185–190.

Feinstein, A., Brown, R., & Ron, M. (1994). Effects of practice of serial tests of attention in healthy subjects. *Journal of Clinical and Experimental Neuropsychology, 16*, 436–447.

Heaton, R. K. (1992). *Comprehensive norms for an expanded Halstead-Reitan Battery: A supplement for the WAIS-R*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive Norms for an Expanded Halstead-Reitan Battery*. Odessa, FL: Psychological Assessment Resources.

Hermann, B. P., Wyler, A. R., VanderZwagg, R., LeBailly, R. K., Whitman, S., Sommes, G., & Ward, J. (1991). Predictors of neuropsychological change following anterior temporal lobecotomy. Role of regression toward the mean. *Epilepsy, 4*, 139–148.

Horton, A. M. (1992). Neuropsychological practice effects related to age: A brief note. *Perceptual & 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.

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

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

Spitzer, R. L., & Williams, J. B. W. (1986). *Structured Clinical Interview for DSM III-R Patient Version (SCID-P, 5186)*. New York: Biometric Research Department of the New York State Psychiatric Institute.

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.

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

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.