Ψ Psychology Press
Taylor & Francis Group

THE PREDICTION OF CHANGE: NORMATIVE NEUROPSYCHOLOGICAL TRAJECTORIES

Deborah K. Attix[1], Tyler J. Story[1], Gordon J. Chelune[2], J. D. Ball[3], Michael L. Stutts[3], Robert P. Hart[4], and Jeffrey T. Barth[5]

[1]*Divisions of Neurology and Medical Psychology, Duke University Medical Center, Durham, NC, USA,* [2]*University of Utah Health Sciences Center, Salt Lake City, UT, USA,* [3]*Eastern Virginia Medical School, Norfolk, VA, USA,* [4]*Virginia Commonwealth University, Richmond, VA, USA, and* [5]*University of Virginia School of Medicine, Charlottesville, VA, USA*

While the application of normative standards is vital to the practice of clinical neuropsychology, data regarding normative change remains scarce despite the frequency of serial assessments. Based on 285 normal individuals, we provide co-normed baseline data with demographic adjustments and test-retest standardized regression based (SRB) models for three time points for several measures. These models delineate normal, expected change across time, and yield standardized z-scores that are comparable across tests. Using a new approach, performance on any previous trial was accounted for in the subsequent models of change, yielding serial normative formulas that model change trajectories rather than simple change from point to point. These equations provide indices of deviation from expected baseline and change for use in clinical or research settings.

INTRODUCTION

Because norm-referenced scores provide the basis of interpretative statements concerning a patient's functional status, diagnosis, and treatment needs, recent texts stress not only the importance of informed test selection but also the importance of selecting norms that are appropriate for the goals of the assessment (Lezak, Howieson, & Loring, 2004; Mitrushina, Boone, Razani, & D'Elia, 2005; Strauss, Sherman, & Spreen, 2006). Analysis of change is central to addressing diagnostic questions, and it is based on a discrepancy model in which an *observed* performance is noted to deviate from an *expected* level (Hawkins & Tulsky, 2003). Heaton and colleagues (Heaton, Taylor, & Manly, 2003) observe that there is overwhelming evidence that demographic factors such as age, education, gender, and ethnicity significantly affect an individual's observed cognitive performance. By adjusting normative data for relevant demographic factors, an individual comparison

standard is generated against which a specific patient's observed scores can be compared (Mitrushina et al., 2005; Strauss et al., 2006). While population-based norms describe where a patient's performance falls relative to the mean of the general population, demographically adjusted norms inform the clinician how far a patient's observed performance deviates from his or her expected performance (Busch, Chelune, & Suchy, 2006). The diagnostic value of these deviations can be evaluated using base-rate data, test-operating characteristics such as sensitivity and specificity, receiver-operating characteristics, and estimates of relative risk (Ivnik et al., 2000, 2001), and lend themselves to evidence-based research and practice (APA Presidential Task Force on Evidence-Based Practice, 2006; Chelune, 2002).

Demographically adjusted norms have been developed for many individual test measures such as the California Verbal Learning Test (Delis, Kramer, Kaplan, & Ober, 2000), Hopkins Verbal Learning Test (Vanderploeg et al., 2000), MicroCog (Powell, Kaplan, Whitla, Catlin, & Funkenstein, 1993), Mattis Dementia Rating Scale (Lucas et al., 1998; Schmidt et al., 1994), Ruff Figural Fluency Test (Ruff, 1996), and Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1993). While such norms provide useful individual comparison standards for evaluating the diagnostic value of the specific test, caution must be exercised when using such tests within the context of a battery since the tests were standardized on different populations at different points in time, and the intercorrelations between measures and the base rate of discrepancies scores are often not known. To overcome these limitations, there is a growing appreciation of the value of test batteries that have been *co-normed* using the same population with known demographic features. Examples of such demographically adjusted norms include those for the expanded Halstead-Reitan Battery (Heaton, Miller, Taylor, & Grant, 2004), the Mayo Older Americans Normative Studies (Ivnik, Malec, Smith, Tangalos, & Petersen, 1996; Ivnik et al., 1990, 1992a, 1992b, 1997; Ivnik, Tangalos, Petersen, Kokmen, & Kurland, 1992c; Lucas et al., 1998; Malec et al., 1992), the third editions of the Wechsler Intelligence and Memory Scales (Taylor & Heaton, 2001; The Psychological Corporation, 2002; Wechsler, 1997), the Neuropsychological Assessment Battery (NAB; Stern & White, 2001), and the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, 1998). The NAB and RBANS are notable because they are among the few test batteries that are intended for both single-point assessments as well as repeated use, albeit using alternative forms.

While neuropsychological tests are generally designed to assess the current state or capacity of the individual, there are an increasing number of situations in which patients require serial testing to monitor change in cognitive status as a function of disease progression, treatment response, surgical or pharmacological intervention, and recovery of function (Busch et al., 2006; Chelune, 2003; Chelune, Naugle, Luders, Sedlak, & Awad, 1993; Collie, Maruff, Darby, & McStephen, 2003; Lineweaver & Chelune, 2003; Strauss et al., 2006). In these situations, the patient's observed baseline performance becomes the individual comparison standard against which test–retest change is evaluated, and the variable of interest is the test–retest discrepancy or change score. Interpretation of the clinical signifi-cance of these change scores is complicated by practice effects, measurement error, and regression to the mean (Bruggemans, Van de Vijver, & Huysmans, 1997;

Chelune, 2003; Lineweaver & Chelune, 2003). Interpretation of change scores at the level of the individual is further complicated by patient characteristics such as age, education, and baseline ability, which can influence rate of change although not necessarily uniformly across individuals. While use of alternate forms has been suggested as a useful approach to minimize practice effects (Benedict & Zgaljardic, 1998; Delis et al., 2000; Randolph, 1998; Stern & White, 2001), research still demonstrates significant practice gains (Beglinger et al., 2005; Hinton-Bayre & Geffen, 2005), as alternative forms do not control for procedural or skill-based learning or other factors that contribute to the overall practice effect (Busch et al., 2006). Because alternate forms do not fully remedy practice effects nor meet the psychometric challenges posed by bias and error inherent in serial assessments, methods of assessing reliable change have been developed.

Reliable change methods are essentially a family of statistical procedures that attempt to take into account with varying degrees of success practice effects, measurement error, and regression to the mean by describing the spread or distribution of test–retest change scores that would be expected to occur in the absence of true change (Basso, Bornstein, & Lang, 1999; Chelune et al., 1993; Dikman, Heaton, Grant, & Timken, 1999; Hermann et al., 1996; Iverson, 2001; Jacobson & Truax, 1991; McSweeny, Naugle, Chelune, & Luders, 1993; Sawrie, Chelune, Naugle, & Luders, 1996). Meaningful change is typically determined when an observed difference score exceeds a specified confidence interval set around the mean of the expected change score. These procedures are increasingly being used in outcomes research such as epilepsy surgery (Chelune et al., 1993; Hermann et al., 1996, 1999; Martin et al., 2002; Sawrie et al., 1996; Seidenberg et al., 1998), cardiac procedures (Andrew, Baker, Bennetts, Kneebone, & Knight, 2001; Bruggemans, van de Vijver, & Huysmans, 1999; Collie, Darby, Falleti, Silbert, & Maruff, 2002; Kneebone, Andrew, Baker, & Knight, 1998; Lehrner et al., 2005; M. S. Lewis, Maruff, Silbert, Evered, & Scott, 2006), traumatic brain injury (Dikman et al., 1999; Ferland, Ramsay, Engeland, & O'Hara, 1998; McCrea et al., 2005; Temkin, Heaton, Grant, & Dikmen, 1999), post-operative cognitive dysfunction (Farag, Chelune, Schubert, & Mascha, 2006; M. Lewis, Maruff, & Silbert, 2004; Maze & Todd, 2007; Murkin, 2001), and aging (Duff et al., 2005; Ivnik et al., 1999; Knight, McMahon, Skeaff, & Green, 2007; Raymond, Hinton-Bayre, Radel, Ray, & Marsh, 2006; Sawrie, Marson, Boothe, & Harrell, 1999; Tombaugh, 2005).

While there is growing interest in the use of longitudinal multivariate mixed models to assess cognitive change (Chu et al., 2007; Salthouse, 2007), the use of Standardized Regression-Based (SRB) models remains one of the most powerful means for assessing reliable change at the level of the individual (Chelune, 2003). Introduced by McSweeny and colleagues (1993), the SRB approach involves regression modeling to derive prediction equations for retest scores based on initial baseline performances. In addition to baseline scores, these regression equations are often multivariate and include demographic variables and other potentially relevant factors that might affect the rate of test–retest change (Chelune, 2003; Hermann et al., 1996; Sawrie et al., 1996). Norms for expected change can be derived in the form of standardized z-scores by dividing the difference between the *observed* and regression *predicted* retest scores by the

Standard Error of the Estimate (SEE) for the regression equation. Similar to demographically adjusted norms for single-point evaluations, the SRB equations can be used with individual patients to determine whether the patient's observed rate of change deviates significantly from expectation. SRB norms of expected change also have the advantage of taking into account differences in reliability and susceptibility to practice between test measures such that change scores for groups of variables are expressed on a common metric, facilitating between-variable discrepancy analysis (Chelune, 2003).

The current study is divided into two parts, and presents normative baseline and test–retest change data over three time points for a sample of normal individuals. Part 1 provides co-normed data with subsequent demographic adjustments for 14 variables derived from eight commonly used neuropsychological tests. Part 2 provides SRB norms of expected change across three time points. In addition to providing the SRB prediction equations to forecast expected change from Time-1 (T_1) to Time-2 (T_2), we also apply a relatively new approach (Chelune, Attix, & Story, 2007; Chelune, Ivnik, & Smith, 2006) for predicting change at Time-3 (T_3). Whereas previous authors (e.g., McCrea et al., 2005) have employed the SRB approach to predict retest performance across multiple time points using baseline scores alone, we used both baseline (T_1) and initial change (T_2–T_1) to predict subsequent change scores. This method models the potential impact of differential practice effects on subsequent retest performances, yielding prediction equations that reflect neuropsychological trajectories over time.

METHOD

Participants

An archival sample of 285 normal participants who took part in an investigation of the potential human health effects of exposure to a naturally occurring dinoflaggellate, *Pfiesteria Piscicida*, was used. The parent study was an epidemiological surveillance research project funded by the Centers for Disease Control and Prevention utilizing a prospective longitudinal design. North Carolina and Virginia participants were assessed for the detection of *Pfiesteria*-related health problems associated with occupational or recreational water contact. While there were limited *Pfiesteria*-associated fish kills during the study period, a series of analyses on variables reflecting various parameters of exposure was conducted. These parent study analyses failed to reveal any consistent, significant association between exposure indicators and neuropsychological test performance (Turf et al., 1999; Moe, 2004).

Participants were aged 18 to 71 ($M = 43.4$, $SD \pm 11.7$; 15% aged 18–30; 82.8% aged 31–64; 2.2% 65+), with an average education of 12.4 years ($SD \pm 2.5$; range 6–20, 23.6% < 12, 39.3% = 12, 37.4% > 12) and a mean Wide Range Achievement Test-3 Reading scaled score of 92.7 ($SD \pm 13.9$). They were primarily male (244 males/41 females), and all were Caucasian. Participants had no known neurological conditions, history of treatment for substance abuse, or diagnosis of psychosis/treatment with antipsychotic agents. Also excluded were participants with a history of traumatic brain injury with loss of consciousness longer than 30 minutes

or persisting cognitive sequelae sufficient to interfere with daily functioning, a history of solvent or pesticide poisoning, insulin-dependent diabetes, or a history of placement in school classes for developmentally disabled individuals. Finally, those who had participated in other studies involving *Pfiesteria* or estuary syndromes were excluded.

Study participants were examined at entry and every 5 to 7 months with a medical exam and neuropsychological assessment over a 1.5-year period. The mean test–retest interval was 177 (*SD* 51) days for Time 2, and 195 (*SD* 91) days for Time 3. In North Carolina participants also completed biweekly telephone interviews to document perceived symptoms. Some participants completed exams "triggered" by exposure to a fishkill or cognitive symptom complaints. In such cases, the subsequent assessment was scheduled for 5–7 months after the triggered exam to maintain an equal exam interval between participants. Performance as a function of routine versus triggered visits did not differ.

The Institutional Review Board of all participating medical centers approved the study (Duke University Medical Center, University of North Carolina-Chapel Hill, Virginia Commonwealth University, and Eastern Virginia Medical School). Participants provided written informed consent, and were compensated $100 for each half-day medical and neuropsychological examination. Only participants with data for all three visits were included in the current analysis for each measure.

Procedures

Table 1 outlines the test instruments used. Standardized administration and scoring procedures were utilized for each neuropsychological measure, with the exception of minor modifications for the Rey-Osterreith Complex Figure Test (ROCFT) and Rey Auditory Verbal Learning Test (RAVLT). Specifically, only copy and delayed recall trials of the ROCFT were given. In order to avoid ceiling effects, administration alternated between two versions of the RAVLT (A-B-A; immediate recall after the interference list was given for all participants, but only a subset of these data was archived and available for analyses from participating sites). The use of regression modeling to control for practice effects and other sources of error was anticipated, and thus alternate forms were only administered when ceiling effects were expected (i.e., RAVLT), and were administered in the same

Table 1   Neuropsychological measures

Wide Range Achievement Test – 3, (baseline only)	Wilkinson, 1993
Symbol Digit Modalities Test, written version	Smith, 1982
Trail Making Test	U.S. Army, 1944
Letter–Number Sequencing, Wechsler Adult Intelligence Scale-III	Wechsler, 1997
Stroop Color and Word Test	Golden, 1978
Controlled Oral Word Association, Multilingual Aphasia Examination	Benton & Hamsher, 1976
Rey Auditory Verbal Learning Test	Schmidt, 1996
Rey-Osterreith Complex Figure Test	Meyers & Meyers, 1995
Grooved Pegboard Test	Klove, 1963

order for all participants, such that covarying test version was unnecessary. Time was also not examined as a performance modifier as the test–retest interval was by study design restricted in range.

Data analyses

Analyses proceeded in two parts: Part 1 – baseline co-norming, and Part 2 – normative change over time. In Part 1, stepwise linear regression was used to create formulas that delineate co-normed demographically corrected baseline performance for the sample, using age, education, and sex as predictors. These equations provide indices of deviation from expected baseline performance for use in clinical or research settings. In Part 2 stepwise linear regression and SRB procedures were utilized to predict normative change trajectories over time, with age, education, sex, and previous raw test scores as potential predictors.

RESULTS

Part 1: Baseline co-normed performance

Mean raw scores on the variables of interest for the sample at baseline (T_1), Time 2 (T_2), and Time 3 (T_3) examinations are presented in Table 2. Significant differences between time points are noted in Table 2 based on one-way repeated measures ANOVAs and post-hoc Tukey tests. Based on these analyses, statistically significant practice effects for each measure were observed over the three time points.

Table 2   Sample means and standard deviations over time

	n	T_1 Mean (*SD*)	T_2 Mean (*SD*)	T_3 Mean (*SD*)	Change statistic	
SDMT	285	46.45 (9.6)	48.36 (10.1)	48.52 (10.3)	✱	☽
Trails A	284	30.90 (10.9)	29.70 (9.6)	28.37 (9.4)	◆	☽
Trails B	285	79.12 (31.6)	71.49 (28.3)	71.71 (30.0)	✱	☽
L–N Sequencing – raw	285	10.31 (2.3)	10.77 (2.5)	10.87 (2.4)	✱	☽
Stroop C/W	278	37.44 (8.7)	40.62 (8.5)	41.68 (9.0)	✱ ◆	☽
COWA	254	33.57 (10.6)	35.85 (11.5)	37.16 (11.9)	✱ ◆	☽
RAVLT Total	234	48.53 (9.0)	46.42 (9.4)	51.05 (9.6)	✱ ◆	☽
RAVLT Short Delay	186	10.26 (2.7)	9.77 (2.8)	11.23 (2.8)	✱ ◆	☽
RAVLT Long Delay	234	9.77 (2.9)	9.17 (3.2)	10.53 (3.2)	✱ ◆	☽
RAVLT Recognition	234	13.59 (1.5)	13.48 (1.6)	13.78 (1.6)	◆	
Rey-O Copy	284	28.00 (4.5)	28.26 (4.2)	28.69 (4.1)		☽
Rey-O Delay	285	15.35 (5.3)	17.41 (5.4)	18.35 (5.6)	✱ ◆	☽
GP Dom	283	77.78 (15.6)	75.73 (14.9)	73.45 (14.3)	✱ ◆	☽
GP NonDom	285	84.87 (16.0)	82.26 (16.1)	79.68 (15.6)	✱ ◆	☽

SDMT = Symbol Digit Modalities Test; L–N Sequencing = Letter–Number Sequencing; Stroop C/W = Stroop Color/Word trial; COWA = Controlled Word Association Test; RAVLT = Rey Auditory Verbal Learning Test; Rey-O = Rey-Osterreith; GP Dom = Grooved Pegboard Dominant Hand; GP NonDom = Grooved Pegboard Nondominant Hand.
T_1 = First Exam; T_2 = Second Exam; T_3 = Third Exam; n = sample size; *SD* = Standard Deviation. ✱ T_1 and T_2 different at $p < .05$. ◆ T_2 and T_3 different at $p < .05$. ☽ T_1 and T_3 different at $p < .05$.

The RAVLT was the only measure showing a decline rather than gain at T_2, likely due to differences in alternate-form difficulty.

Normative performance was estimated using linear regression with all data available for the sample. Age, education, and sex were entered into all models as potential predictors. Normative formulas based on the models are found in Table 3. These models account for 5–45% of the variance in the sample as noted by the R^2 values in Table 2.

Application. For use of these formulas in a clinical setting, consider the following case example. A 56-year-old male with 10 years of education obtains a baseline score of 32 on the Stroop Test. This score falls 5.44 points below the sample mean of 37.44 (see Table 2), or -0.63 SD (z-scores) below average ($z = -5.44/8.7 = -0.63$). While we know that this individual falls -0.63 SD below the normative sample's mean or at the 27th percentile, we do not know whether this person's performance actually deviates from other men with similar demographic characteristics. However, if we apply the regression-based formulas in Table 3 we can more accurately delineate his actual divergence from his demographic cohort by considering the relevant modifying demographic variables, which account for close to half of the variance of some measures. The patient's predicted score is first calculated using the regression equation in Table 3 for Stroop, which is then subtracted from the observed score and divided by the standard error of the estimate (SEE).

$$\text{Predicted Stroop} = (31.37 + (-.18 \times \text{age}) + (1.12 \times \text{education})) = 32.49$$

$$\text{Deviation from Expected Score} = \frac{\text{Observed score}(32) - \text{Predicted score}(32.49)}{\text{Standard Error of the Estimate}(8.04)} = \text{z-score of} - 0.06$$

Table 3 Formulas predicting baseline performance

	SEE		R^2
SDMT	7.18	$= 41.31 + (\text{Age} \times -.30) + (\text{Education} \times 1.84) + (\text{Sex}^* \times -5.66)$	45%
Trails A	10.35	$= 25.78 + (\text{Age} \times .20) + (\text{Education} \times -.60) + (\text{Sex}^* \times 4.35)$	10%
Trails B	28.41	$= 89.10 + (\text{Age} \times .68) + (\text{Education} \times -3.99) + (\text{Sex}^* \times 11.77)$	20%
L–N Seq	2.13	$= 6.72 + (\text{Age} \times -.02) + (\text{Education} \times .37)$	16%
Stroop C/W	8.04	$= 31.37 + (\text{Age} \times -.18) + (\text{Education} \times 1.12)$	15%
COWA	10.12	$= 17.74 + (\text{Education} \times 1.31)$	9%
RAVLT Total	7.90	$= 51.14 + (\text{Age} \times -.23) + (\text{Education} \times 1.05) + (\text{Sex}^* \times -5.98)$	23%
RAVLT Short Delay	2.54	$= 10.75 + (\text{Age} \times -.04) + (\text{Education} \times .27) + (\text{Sex}^* \times -2.32)$	15%
RAVLT Long Delay	2.65	$= 12.06 + (\text{Age} \times -.06) + (\text{Education} \times .20) + (\text{Sex}^* \times -2.64)$	19%
RAVLT Recognition	1.50	$= 14.47 + (\text{Sex}^* \times -1.01)$	5%
Rey-O Copy	4.33	$= 25.32 + (\text{Age} \times -.07) + (\text{Education} \times .46)$	10%
Rey-O Delay	5.08	$= 14.53 + (\text{Age} \times -.11) + (\text{Education} \times .46)$	10%
GP Dom	14.08	$= 62.14 + (\text{Age} \times .40) + (\text{Education} \times -.84) + (\text{Sex}^* \times 10.32)$	19%
GP NonDom	15.34	$= 71.96 + (\text{Age} \times .45) + (\text{Education} \times -1.18) + (\text{Sex}^* \times 9.57)$	19%

SDMT = Symbol Digit Modalities Test; L–N Sequencing = Letter–Number Sequencing; Stroop C/W = Stroop Color/Word trial; COWA = Controlled Word Association Test; RAVLT = Rey Auditory Verbal Learning Test; Rey-O = Rey-Osterreith; GP Dom = Grooved Pegboard Dominant Hand; GP NonDom = Grooved Pegboard Nondominant Hand.
SEE = Standard Error of the Estimate; R^2 = Variance accounted for. *Sex: Male = 1; Female = 0.

In this example we see that the patient's baseline score of 32, while more than half *SD* below the average of the general normative sample ($z = -0.63$), actually falls near expectation ($z = -0.06$) relative to his demographic peers (48th percentile).

Part 2: Trajectories of neuropsychological change

Mean changes on the variables of interest for the sample across time were calculated by subtracting the previous raw score from the relevant subsequent raw score (e.g., T_2–T_1 for calculation of average change from $T_1 \rightarrow T_2$). The means and *SD*s of the change scores are presented in Table 4.

Table 4   Sample change means and standard deviations

	n	Change $T_1 \rightarrow T_2$ Mean (*SD*)	Change $T_2 \rightarrow T_3$ Mean (*SD*)	Change $T_1 \rightarrow T_3$ Mean (*SD*)
SDMT	285	1.91 (5.7)	0.16 (6.0)	2.07 (6.0)
Trails A	284	−1.20 (10.4)	−1.33 (8.7)	−2.53 (10.8)
Trails B	285	−7.62 (24.4)	0.21 (25.6)	−7.41 (25.5)
L–N Seq – raw	285	0.46 (2.1)	0.11 (2.1)	0.56 (2.1)
Stroop C/W	278	3.18 (6.2)	1.06 (6.1)	4.25 (6.5)
COWA	254	2.28 (6.4)	1.31 (6.8)	3.59 (7.3)
RAVLT Total	234	−2.11 (8.3)	4.63 (8.9)	2.53(7.6)
RAVLT Short Delay	186	−0.49 (2.6)	1.46 (2.6)	0.97 (2.3)
RAVLT Long Delay	234	−0.60 (2.6)	1.36 (2.7)	0.76 (2.4)
RAVLT Recognition	234	−0.12 (1.8)	0.30 (1.6)	0.19 (1.6)
Rey-O Copy	284	0.25 (4.3)	0.43 (3.6)	0.69 (4.3)
Rey-O Delay	285	2.06 (4.4)	0.94 (4.4)	3.01 (4.6)
GP Dom	283	−2.05 (10.6)	−2.28 (9.9)	−4.33 (10.7)
GP NonDom	285	−2.62 (11.8)	−2.57 (11.1)	−5.19 (10.3)

SDMT = Symbol Digit Modalities Test; L–N Sequencing = Letter–Number Sequencing; Stroop C/W = Stroop Color/Word trial; COWA = Controlled Word Association Test; RAVLT = Rey Auditory Verbal Learning Test; Rey-O = Rey-Osterreith; GP Dom = Grooved Pegboard Dominant Hand; GP NonDom = Grooved Pegboard Nondominant Hand.
T_1 = First Exam; T_2 = Second Exam; T_3 = Third Exam; *n* = sample size; *SD* = Standard Deviation.
These change scores are average raw scores for the entire sample and do not consider patient characteristics, change score modifiers, or how these interact. These are the average change scores, and thus reflect the averaging of the effects of the predictors.

Stepwise linear regression and SRB methods were then utilized to model change over time. Age, education, and sex were entered into all models as potential predictors of change. Baseline performance and change across previous trials when applicable was also entered into each model (i.e., baseline raw score entered into models for $T_1 \rightarrow T_2$ change, $T_2 \rightarrow T_3$ change, and $T_1 \rightarrow T_3$ change; $T_1 \rightarrow T_2$ change score also entered into models predicting $T_2 \rightarrow T_3$ change and $T_1 \rightarrow T_3$ change). Raw scores and change scores were used rather than standardized scores in these models, along with potential demographic predictors. Table 5 presents the resulting regression formulas for predicting the amount of expected change for each measure; note that previous performance consistently emerged as a significant predictor of subsequent performance.

Table 5 Prediction of change formulas

	R^2	SEE	Predicted Δ:
SDMT			
$\Delta T_1 \rightarrow T_2$	5%	5.59	$= 7.82 + (T1 \times -.13)$
$\Delta T_2 \rightarrow T_3$	23%	5.30	$= 1.12 + (\Delta_{1-2} \times -.50)$
$\Delta T_1 \rightarrow T_3$	23%	5.30	$= 1.12 + (\Delta_{1-2} \times .50)$
Trails A			
$\Delta T_1 \rightarrow T_2$	39%	8.12	$= 18.53 + (T1 \times -.62) + (Age \times .14) + (Education \times -.53)$
$\Delta T_2 \rightarrow T_3$	30%	7.36	$= 10.83 + (T1 \times -.43)^{\ddot{y}\,\ddot{y}} + (\Delta_{1-2} \times -.58) + (Age + .13) + (Education \times -.41)$
$\Delta T_1 \rightarrow T_3$	54%	7.36	$= 10.83 + (T1 \times -.43) + (\Delta_{1-2} \times .42) + (Age \times .13) + (Education \times -.41)$
Trails B			
$\Delta T_1 \rightarrow T_2$	33%	20.20	$= 28.10 + (T1 \times -.49) + (Age \times .34) + (Education \times -1.55) + (Sex^* \times 8.57)$
$\Delta T_2 \rightarrow T_3$	34%	21.03	$= 20.38 \ (T1 \times -.34) + (\Delta_{1-2} \times -.73) + (Age \times .41) + (Education \times -1.33)$
$\Delta T_1 \rightarrow T_3$	33%	21.03	$= 20.38 + (T1 \times -.34) + (\Delta_{1-2} \times .27) + (Age \times .41) + (Education \times -1.33)$
L–N Sequencing			
$\Delta T_1 \rightarrow T_2$	18%	1.92	$= 2.11 + (T1 \times -.41) + (Education \times .21)$
$\Delta T_2 \rightarrow T_3$	31%	1.73	$= 3.66 + (T1 \times -.23) + (\Delta_{1-2} \times -.57) + (Age \times -.02)$
$\Delta T_1 \rightarrow T_3$	33%	1.73	$= 3.66 + (T1 \times -.23) + (\Delta_{1-2} \times .43) + (Age \times -.02)$
Stroop C/W			
$\Delta T_1 \rightarrow T_2$	19%	5.63	$= 12.58 + (T1 \times -.32) + (Education \times .38) + (Sex^* \times -2.59)$
$\Delta T_2 \rightarrow T_3$	26%	5.34	$= 8.56 + (T1 \times -.16) + (\Delta_{1-2} \times -.52) + (Age \times -.08) + (Education \times .32)$
$\Delta T_1 \rightarrow T_3$	33%	5.34	$= 8.56 + (T1 \times -.16) + (\Delta_{1-2} \times .48) + (Age \times -.08) + (Education \times .32)$
COWA			
$\Delta T_1 \rightarrow T_2$	2%	6.36	$= 5.45 + (T1 \times -.09)$
$\Delta T_2 \rightarrow T_3$	16%	6.26	$= 2.27 + (\Delta_{1-2} \times -.42)$
$\Delta T_1 \rightarrow T_3$	26%	6.26	$= 2.27 + (\Delta_{1-2} \times .58)$
RAVLT Total			
$\Delta T_1 \rightarrow T_2$	16%	7.60	$= 16.03 + (T1 \times -.37)$
$\Delta T_2 \rightarrow T_3$	41%	6.91	$= 12.37 + (T1 \times -.19) + (\Delta_{1-2} \times -.74)$
$\Delta T_1 \rightarrow T_3$	18%	6.91	$= 12.37 + (T1 \times -.19) + (\Delta_{1-2} \times .26)$
RAVLT Short Delay			
$\Delta T_1 \rightarrow T_2$	19%	2.38	$= 3.91 + (T1 \times -.43)$
$\Delta T_2 \rightarrow T_3$	42%	1.99	$= 3.35 + (T1 \times -.22) + (\Delta_{1-2} \times -.70)$
$\Delta T_1 \rightarrow T_3$	26%	1.99	$= 3.35 + (T1 \times -.22) + (\Delta_{1-2} \times .30)$
RAVLT Long Delay			
$\Delta T_1 \rightarrow T_2$	12%	2.45	$= 2.41 + (T1 \times -.31)$
$\Delta T_2 \rightarrow T_3$	38%	2.17	$= 2.54 + (T1 \times -.16) + (\Delta_{1-2} \times -.68)$
$\Delta T_1 \rightarrow T_3$	20%	2.17	$= 2.54 + (T1 \times -.16) + (\Delta_{1-2} \times .32)$
RAVLT Recognition			
$\Delta T_1 \rightarrow T_2$	27%	1.51	$= 8.06 + (T1 \times -.60)$
$\Delta T_2 \rightarrow T_3$	39%	1.27	$= 5.06 + (T1 \times -.26) + (\Delta_{1-2} \times -.62) + (Education \times -.11)$
$\Delta T_1 \rightarrow T_3$	38%	1.27	$= 5.06 + (T1 \times -.26) + (\Delta_{1-2} \times .38) + (Education \times -.11)$
Rey-O Copy			
$\Delta T_1 \rightarrow T_2$	31%	3.58	$= 16.95 + (T1 \times -.54) + (Age \times -.04)$
$\Delta T_2 \rightarrow T_3$	29%	3.06	$= 7.79 + (T1 \times -.34) + (\Delta_{1-2} \times -.54) + (Education \times .18)$
$\Delta T_1 \rightarrow T_3$	50%	3.06	$= 7.79 + (T1 \times -.34) + (\Delta_{1-2} \times .46) + (Education \times .18)$
Rey-O Delay			
$\Delta T_1 \rightarrow T_2$	22%	3.91	$= 6.98 + (T1 \times -.39) + (Age \times -.07) + (Education \times .32)$
$\Delta T_2 \rightarrow T_3$	22%	3.85	$= 4.42 + (T1 \times -.16) + (\Delta_{1-2} \times -.51)$
$\Delta T_1 \rightarrow T_3$	41%	8.30	$= 8.75 + (T1 \times -.23) + (\Delta_{1-2} \times .44) + (Age \times .13)$
GP NonDom			
$\Delta T_1 \rightarrow T_2$	19%	10.67	$= 19.36 + (T1 \times -.33) + (Age \times .13)$
$\Delta T_2 \rightarrow T_3$	45%	8.30	$= 12.20 + (T1 \times -.23) + (\Delta_{1-2} \times -.69) + (Age \times .18) + (Education \times -.42)$
$\Delta T_1 \rightarrow T_3$	36%	8.30	$= 12.20 + (T1 \times -.23) + (\Delta_{1-2} \times .31) + (Age \times .18) + (Education \times -.42)$

SDMT = Symbol Digit Modalities Test; L–N Sequencing = Letter–Number Sequencing; Stroop C/W = Stroop Color/Word trial; COWA = Controlled Word Association Test; RAVLT = Rey Auditory Verbal Learning Test; Rey-O = Rey-Osterreith; GP Dom = Grooved Pegboard Dominant Hand; GP NonDom = Grooved Pegboard Nondominant Hand. SEE = Standard Error of the Estimate; R^2 = Variance accounted for. Δ = Change. T_1 = First Exam; Δ_{1-2} = Change from T_1 to T_2; T3 = Third Exam; n = sample size; SD = Standard Deviation. *Sex: Male = 1; Female = 0.

Application. By performing simple computations, the clinician can use the SRB formulas presented in Part 2 to determine if a patient's observed change or practice effect deviates from expectation based on data from our normative sample. Rather than relying on accepted tradition or clinical judgment, these formulas provide an objective estimate of expected variability across multiple testing sessions while including the effects of variables that modify change. Consider the previous clinical example.

Our 56-year-old male obtained a score of 32 at baseline, but subsequently obtained scores of 23 and 25. In terms of reported events, the family described a change in cognition after he returned from vacation, but he himself was not aware of any precipitous event or cognitive decline. This prompted the second evaluation. The patient's second score of 23 represents a 9-point decrement from baseline, with the difference between the first and second score (32–23) falling at about 1 SD of the baseline sample mean ($SD = 8.7$; see Table 2), which is an often applied threshold for clinically significant decline. By this standard, the observed change is a matter of concern but is not unequivocal evidence of clinically significant change. To assess whether this decrement from baseline is truly meaningful, we must consider the distribution of normal change scores and account for the potential impact demographic variables have on this distribution.

When considering average change across time in Table 4, it appears that the patient's first decline differs considerably from the 3-point gain observed in this sample. However, the SD in change scores indicates considerable variability in change scores within this sample. To evaluate the significance of this change while considering the impact of performance and demographic variables, we apply the SRBs. We begin by estimating the patient's predicted change score using the Stroop regression equation for $\Delta T_1 \rightarrow T_2$ in Table 5. We then subtract this score from the observed $T_1 \rightarrow T_2$ difference and divide by the SEE.

$$\text{Predicted } \Delta T_1 \rightarrow T_2 = 12.58 + (32 \times -.32) + (10 \times .38) + (1 \times -2.59) = 3.55$$

$$\text{Deviation from Expected } \Delta T_1 \rightarrow \Delta T_2 = \frac{\text{Observed } \Delta (-9) - \text{Predicted } \Delta (3.55)}{\text{Standard Error of the Estimate} (5.63)} = (z = -2.23)$$

According to this outcome, the 9-point decline for this patient is indeed statistically rare, and likely to occur in only 1.3% of the normative sample. Given these data, which consider the patient's background as well as sources of error and bias, we no longer have a questionable change but rather a decline that clearly deviates from expectation.

The utility of the SRB method is also illustrated when considering this individual's change in performance by Time 3. His 2-point gain (23 to 25) actually falls above the average T2 to T3 change ($M = 1.06$, $SD = 6.1$; see Table 4), which at first glance might be considered an improved performance and consistent with expectation. When using the $\Delta T_2 \rightarrow T_3$ SRB equation, however, we see that this improvement actually deviates from his expected change by a z score of $-.91$ because his baseline performance, the prior T_1 to T_2 change, and demographic variables are considered. The use of these variables in the SRB model yields a more precise assessment of his expected trajectory of change. A similar point is made when considering the overall 7-point decrement from Time 1 to Time 3 (32–25).

When compared to group average change scores ($M = 4.25$, $SD = 6.5$; see Table 4) the patient's decline falls -1.73 standard deviations below the average practice effect. Use of the $\Delta T_1 \rightarrow T_3$ SRB equation, however, indicates that this overall decrement falls just short of one standard deviation below expected change when considering all relevant variables ($z = -.91$). In each of these situations we see that the SRB equation was particularly useful when considering change over multiple time points and provided more accuracy than comparing performance to group average practice effects using change score means and standard deviations.

DISCUSSION

The practice of clinical neuropsychology is driven by the systematic application of measurement tools and principles, and typically boasts sound and strong normative resources (e.g., Lezak et al., 2004; Mitrushina et al., 2005; Strauss et al., 2006). Yet the amount of data regarding normative *change* remains scarce given the frequency with which the clinician or researcher employs serial assessment for elucidating diagnosis, monitoring intervention outcomes, and tracking disease progression or recovery of function (Chelune, 2002; Chu et al., in press; McCrea et al., 2005). The lack of published data on normative change places the clinician in a position of estimating practice effect changes for each measure used, while considering potential patient modifiers that may impact change as well, such as age, education, and sex. Regression to the mean, measurement error, and the effect of alternate forms must also be worked into the change gestalt (Lewis et al., 2006). Longitudinal studies of normal samples offer us the opportunity to replace educated estimates with concrete data on normative change. Simpler concepts of ''back to baseline'' or ''one *SD* change'' based on the *SD* of the test rather than of change scores are replaced by data illustrating the sophisticated relationship between test scores and covariates over time. In addition to providing co-normed baseline data, this study presents formulas allowing for calculation of deviation from normal longitudinal change across three time points.

The co-standardization of baseline performance measures with appropriate demographic corrections on the same population has the potential to favorably facilitate interpretation and conceptualization and hence is presented for this sample. Co-norming reduces the need for clinically based interpretive adjustments made to correct for differences in norm group demographics, and methods of data standardization or data presentation.

In a similar fashion, an important advantage of using SRB models to construct change norms is that deviations from expected gains are expressed in a standardized *z*-score unit that is comparable across measures. Like co-standardized baseline data, co-normed change formulas yielding a consistent unit of deviation from change enhances interpretation of change, as modeling the norms using the same methodology within the same population essentially levels off the differential effects of practice, demographic modifiers, statistical artifacts, and inconsistent reliabilities across measures.

The increasing attention to the need for change data based on normals has produced SRB model norms across two or more time points (e.g., Duff et al., 2005, Ivnik et al., 1999, Sawrie et al., 1995) for select measures. This study

augments these pioneering efforts by utilizing a relatively new approach wherein the normative data considers baseline and previous change in the models of subsequent change, yielding change estimates that are best conceptualized as reflecting change trajectories.

This study demonstrates important principles of neuropsychological *change* over time. First, while it is well accepted that patient demographics modify cognitive performance (Heaton et al., 2003), review of age, education, and sex variables in these SRB models clearly demonstrate that these variables also often independently modify *change* in performance. Second, modifiers of *baseline* performance versus *change* performance are not always the same. Third, in terms of variables that most consistently modify change, the adage that "...the best predictor of future behavior is past behavior" fits well. Baseline and/or previous change *always* predicted subsequent change. While age and education often emerged as significant predictors, the impact of sex on change was minimal, and no demographic predicted change like previous performance did.

The effect of including baseline performance and previous change when applicable warrants careful attention. Baseline performance is considered in the estimates of change, which is critical, as evidenced by the predictive value of these prior performances in the change models. Low versus high performers will change differentially over time. However, inclusion of these scores in estimates of change also raises important interpretative considerations. Specifically, the degree of *change* from first to second assessment, through inclusion of these scores in the $T_2 \rightarrow T_3$ and $T_1 \rightarrow T_3$ models, will be considered in the predicted change scores for these latter models. Therefore, if the first change is notably discrepant from average, this discrepancy will be worked into the expectation for the change in the next interval. Inaccurate or aberrant baseline scores will also almost always influence the estimate of change at all time points. Nonetheless, inclusion of previous performance allows for consideration of known performance data that is clearly related to subsequent scores. As such, it is most useful to consider the serial normative formulas as truly modeling *trajectories* rather than simple *change from point to point*.

Given that change scores have their own unique distributions, application of these formulas, which model the effects of demographic and performance variables on change, will be most useful when the patient differs from the group average on these variables. Thus, at a minimum, use of SRB models is advocated over computations using group change means and standard deviations whenever the patient's demographics or performance variables deviate considerably from the norm group averages. SRB models also consider the sophisticated relationship among modifying variables and time, and should thus be utilized when estimating change across several time points. And finally, SRB models provide a level of precision that may be especially appropriate for research settings. In all of these cases, use of SRBs will enhance accuracy and precision.

Application of SRB methods to research programs can augment inferential statistics that show group differences that may be reliable but not necessarily meaningful by providing data on base rates of normal change (Chelune, 2002). For instance, a recent study expanded on the inferential analysis of group differences in post-operative cognitive dysfunction between a high versus low depth of anesthesia

study by incorporating SRB-based risk calculations. Using normal change base rates derived from a normal control sample, the investigators compared the percentage of patients in each anesthesia condition whose change scores exceeded those expected by chance in the normal sample (Farag et al., 2006).

It is important to note that because this study used a normal sample, the models are most useful for determining whether scores fall within the range of normal individuals. Rather than addressing abnormality, these models provide information about the rarity of change scores in a normal population. In contrast to those built on normal individuals, SRB models built on patient groups would provide valuable information about the course of illness, which can be important in gauging deviation from normal course or the effect of treatment. Indeed, models based on patient populations would be particularly useful for illnesses involving a dynamic course, such as the progressive deficits of Alzheimer's disease or the resolving deficits of traumatic brain injury. In future studies involving such populations, it will be important to model *change trajectories*. As noted, our application example highlights the effect of a markedly discrepant change score during the first interval on the subsequent expected trajectories and deviation scores. Clinically based SRB equations would be particularly useful in such cases, as the initial decline indicates that comparison to a relevant clinical population would be most valid for subsequent estimates. Also, it is important to emphasize that these data will be most applicable to samples similar to the archival one used here, and would not necessarily generalize to populations that differ considerably. Because ethnic minorities, women, and elders are either not included or are inadequately represented in this sample, application of these norms to these groups should be undertaken with considerable caution. These models may also not be applicable for patients seen at considerably shorter or longer test–retest intervals than those of the present sample. Finally, as with baseline norms, change norms require use of the same test versions as used in the standardization sample.

While performance norms can be found in multiple texts (e.g., Heaton et al., 2004; Lezak et al., 2004; Mitrushina et al., 2005; Strauss et al., 2006), the paucity of serial normative data available to the clinician represents a pressing problem for our field. The challenge facing clinicians in the absence of such change data is considerable; estimating the differential impact of all the relevant variables accurately across measures and across time is a tremendous challenge. In addition, the effects of demographics, error, and bias on change scores for serial assessments are less understood than the effects of these variables at baseline. The present study was designed to contribute to the evolving norms and methodologies seeking to resolve these normative concerns. In addition to providing co-normed baseline data, these data augment the growing literature on calculating clinically significant, reliable change by providing averages and SRB models for multiple measures over multiple time points, and by modeling previous change into subsequent change formulas, yielding normative change trajectories.

ACKNOWLEDGMENTS

REFERENCES

U.S. Army. (1944). *Army Individual Test Battery*. Washington, DC: U.S. Army.

Andrew, M. J., Baker, R. A., Bennetts, J., Kneebone, A. C., & Knight, J. L. (2001). A comparison of neuropsychologic deficits after extracardiac and intracaradiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia*, *15*(1), 9–14.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285.

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist*, *13*(3), 283–292.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*, *20*(4), 517–529.

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*, *20*, 339–352.

Benton, A., Hamsher, K., & Sivan, A. (1978). *Multilingual Aphasia Examination*. Iowa City, IA: AJA Associates, Inc.

Bruggemans, E. F., Van de Vijver, F. J., & Huysmans, H. A. (1997). Assessment of cognitive deterioration in individual patients following cardiac surgery: Correcting for measurement error and practice effects. *Journal of Clinical and Experimental Neuropsychology*, *19*(4), 543–559.

Bruggemans, E. F., van de Vijver, F. J., & Huysmans, H. A. (1999). Defining neuropsychological deterioration after cardiac surgery. *Annals of Thoracic Surgery*, *67*(1), 297–298.

Busch, R. M., Chelune, G. J., & Suchy, Y. (2006). Using norms in neuropsychological assessment of the elderly. In D. K. Attix & K. A. Welsh-Bohmer (Eds.), *Geriatric neuropsychology: Assessment and intervention* (pp. 133–157). New York: Guilford Press.

Chelune, G. J. (2002). Making neuropsychological outcomes research consumer friendly: A commentary on Keith et al. (2002). *Neuropsychology*, *16*(3), 422–425.

Chelune, G. J. (2003). Assessing reliable neuropsychological change. In R. Franklin (Ed.), *Prediction in forensic and neuropsychology: New approaches to psychometrically sound assessment*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Chelune, G. J., Attix, D., & Story, T. (2007). How reliable are reliable change methods across multiple time points? *Journal of the International Neuropsychological Society*, *13*(S1), 107.

Chelune, G. J., Ivnik, R., & Smith, G. (2006). Application of reliable change methods for identifying abnormal rates of cognitive decline in dementia. *Alzheimer's & Dementia*, *2*(S1), 374.

Chelune, G. J., Naugle, R. I., Luders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52.

Chu, B., Millis, S. R., Arango-Lasprilla, J., Hanks, R., Novack, T., & Hart, T. (2007). Measuring recovery in new learning and memory following traumatic brain injury: A mixed effects modeling approach. *Journal of Clinical and Experimental Neuropsychology*, *29*, 617–625.

Collie, A., Darby, D. G., Falleti, M. G., Silbert, B. S., & Maruff, P. (2002). Determining the extent of cognitive change after coronary surgery: A review of statistical procedures. *Annals of Thoracic Surgery*, *73*(6), 2005–2011.

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*, *9*(3), 419–428.

Delis, D. C., Kramer, J. H., Kaplan, E. F., & Ober, B. A. (2000). *California Verbal Learning Test–Second edition*. San Antonio, TX: The Psychological Corporation.

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

Duff, K., Schoenberg, M. R., Patton, D., Paulsen, J. S., Bayless, J. D., Mold, J., et al. (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology*, *20*(3), 281–290.

Farag, E., Chelune, G. J., Schubert, A., & Mascha, E. J. (2006). Is depth of anesthesia, as assessed by the Bispectral Index, related to postoperative cognitive dysfunction and recovery? *Anesthesia and Analgesia*, *103*(3), 633–640.

Ferland, M. B., Ramsay, J., Engeland, C., & O'Hara, P. (1998). Comparison of the performance of normal individuals and survivors of traumatic brain injury on repeat administrations of the Wisconsin Card Sorting Test. *Journal of Clinical and Experimental Neuropsychology*, *20*(4), 473–482.

Golden, C. J. (1978). *Stroop color and word test: A manual for clinical and experimental uses*. Wood Dale, IL: Stoetling.

Hawkins, K. A., & Tulsky, D. S. (2003). WAIS-III WMS-III discrepancy analysis: Six-factor model index discrepancy base rates, implications, and a preliminary consideration of utility. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik & R. Bornstein, et al. (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 211–272). New York: Academic Press.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test Manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Miller, S. M., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Taylor, M. J., & Manly, J. (2003). Demographic effects and use of demographically corrected norms with the WAIS-III and WMS-III. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik & R. Bornstein, et al. (Eds.), *Clinical interpretation of the WAIS-III and WMS-III*. New York: Academic Press.

Hermann, B. P., Perrine, K., Chelune, G. J., Barr, W., Loring, D. W., Strauss, E., et al. (1999). Visual confrontation naming following left anterior temporal lobectomy: A comparison of surgical approaches. *Neuropsychology*, *13*(1), 3–9.

Hermann, B. P., Seidenberg, M., Schoenfeld, J., Peterson, J., Leveroni, C., & Wyler, A. R. (1996). Empirical techniques for determining the reliability, magnitude, and pattern of neuropsychological change after epilepsy surgery. *Epilepsia*, *37*(10), 942–950.

Hinton-Bayre, A., & Geffen, G. (2005). Comparability, reliability, and practice effects on alternate forms of the Digit Symbol Substitution and Symbol Digit Modalities tests. *Psychological Assessment*, *17*(2), 237–241.

Iverson, G. L. (2001). Interpreting change on the WAIS III/WMS III in clinical samples. *Archives of Clinical Neuropsychology*, *16*, 183–191.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., & Petersen, R. C. (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE Token, WRAT-R Reading, AMNART, STROOP, TMT, and JLO. *The Clinical Neuropsychologist*, *10*(3), 262–278.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., et al. (1992a). Mayo's Older Americans Normative Studies: Updated AVLT norms for ages 56 to 97. *The Clinical Neuropsychologist*, *6*(Suppl.), 83–104.

Ivnik, R. J., Malec, J. F., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1990). The Auditory-Verbal Learning Test (AVLT): Norms for ages 55 years and older. *Psychological Assessment*, *2*, 304–312.

Ivnik, R. J., Malec, J. F., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992b). Mayo's Older Americans Normative Studies: WMS-R norms for ages 56 to 94. *The Clinical Neuropsychologist*, *6*(Suppl.), 49–82.

Ivnik, R. J., Smith, G. E., Cerhan, J. H., Boeve, B. F., Tangalos, E. G., & Petersen, R. C. (2001). Understanding the diagnostic capabilities of cognitive tests. *The Clinical Neuropsychologist*, *15*(1), 114–124.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Petersen, R. C., Boeve, B. F., Kokmen, E., et al. (1999). Testing normal older people three or four times at 1- to 2-year intervals: Defining normal variance. *Neuropsychology*, *13*(1), 121–127.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Tangalos, E. G., Petersen, R. C., & Kokmen, E. (1997). Free and Cued Selective Reminding Test: MOANS norms. *Journal of Clinical & Experimental Neuropsychology*, *19*, 676–691.

Ivnik, R. J., Smith, G. E., Petersen, R. C., Boeve, B. F., Kokmen, E., & Tangalos, E. G. (2000). Diagnostic accuracy of four approaches to interpreting neuropsychological test data. *Neuropsychology*, *14*(2), 163–177.

Ivnik, R. J., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992c). Mayo's Older Americans Normative Studies: WAIS-R norms for ages 56 to 97. *The Clinical Neuropsychologist*, *6*(Suppl.), 1–30.

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.

Klove, H. (1963). Clinical neuropsychology. In F. Forster (Ed.), *The medical clinics of North America*. New York: Saunders.

Kneebone, A. C., Andrew, M. J., Baker, R. A., & Knight, J. L. (1998). Neuropsychologic changes after coronary artery bypass grafting: use of reliable change indices. *Annals of Thoracic Surgery*, *65*(5), 1320–1325.

Knight, R. G., McMahon, J., Skeaff, C. M., & Green, T. J. (2007). Reliable Change Index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Archives of Clinical Neuropsychology*, *22*(4), 513–518.

Lehrner, J., Willfort, A., Mlekusch, I., Guttmann, G., Minar, E., Ahmadi, R., et al. (2005). Neuropsychological outcome 6 months after unilateral carotid stenting. *Journal of Clinical and Experimental Neuropsychology*, *27*(7), 859–866.

Lewis, M., Maruff, P., & Silbert, B. (2004). Statistical and conceptual issues in defining post-operative cognitive dysfunction. *Neuroscience and Biobehavioral Reviews*, *28*(4), 433–440.

Lewis, M. S., Maruff, P., Silbert, B. S., Evered, L. A., & Scott, D. A. (2006). The sensitivity and specificity of three common statistical rules for the classification of post-operative cognitive dysfunction following coronary artery bypass graft surgery. *Acta Anaesthesiol Scand*, *50*(1), 50–57.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment,* (4th ed.). New York: Oxford University Press.

Lineweaver, T. T., & Chelune, G. J. (2003). Use of the WAIS-III and WMS-III in the context of serial assessments: Interpreting reliable and meaningful change. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik & R. Bornstein, et al. (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 303–337). New York: Academic Press.

Lucas, J. A., Ivnik, R. J., Smith, G. E., Bohac, D. L., Tangalos, E.G., Kokmen, E., et al. (1998). Normative data for the Mattis Dementia Rating Scale. *Journal of Clinical & Experimental Neuropsychology*, *20*(4), 536–547.

Malec, J. F., Ivnik, R. J., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., et al. (1992). Mayo's Older Americans Normative Studies: Utility of corrections for age and education for the WAIS-R. *The Clinical Neuropsychologist*, *6*(Suppl.), 31–47.

Martin, R., Sawrie, S., Gilliam, F., Mackey, M., Faught, E., Knowlton, R., et al. (2002). Determining reliable cognitive change after epilepsy surgery: development of reliable change indices and standardized regression-based change norms for the WMS-III and WAIS-III. *Epilepsia*, *43*(12), 1551–1558.

Maze, M., & Todd, M. M. (2007). Special issue on postoperative cognitive dysfunction: Selected reports from the journal-sponsored symposium. *Anesthesiology*, *106*(3), 418–420.

McCrea, M., Barr, W. B., Guskiewicz, K., Randolph, C., Marshall, S. W., Cantu, R., et al. (2005). Standard regression-based methods for measuring recovery after sport-related concussion. *Journal of the International Neuropsychological Society*, *11*(1), 58–69.

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.

Meyers, J., & Meyers, K. (1995). *Rey Complex Figure Test and Recognition Trial: Professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Mitrushina, M. N., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Moe, C. (2004). *Final report to the NC SEARCH Scientific Review Panel*. Unpublished report.

Murkin, J. M. (2001). Editorial: Perioperative neuropsychologic testing. *Journal of Cardiothoracic and Vascular Anesthesia*, *15*, 1–3.

Powell, D. H., Kaplan, E. F., Whitla, D., Catlin, R., & Funkenstein, H. H. (1993). *MicroCog: Assessment of cognitive functioning*. San Antonio, TX: The Psychological Corporation.

Randolph, C. (1998). *Repeatable Battery for the Assessment of Neuropsychological Status: Manual*. San Antonio, TX: The Psychological Corporation.

Raymond, P. D., Hinton-Bayre, A. D., Radel, M., Ray, M. J., & Marsh, N. A. (2006). Test–retest norms and reliable change indices for the MicroCog Battery in a healthy community population over 50 years of age. *The Clinical Neuropsychologist*, *20*(2), 261–270.

Ruff, R. (1996). *Ruff Figural Fluency Test*. Odessa, FL: Psychological Assessment Resources.

Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology*, *21*(4), 401–411.

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

Sawrie, S. M., Marson, D. C., Boothe, A. L., & Harrell, L. E. (1999). A method for assessing clinically relevant individual cognitive change in older populations. *Journal of Gerontology: Psychological Sciences*, *54B*, 116–124.

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A handbook*. Los Angeles, CA: Western Psychological Services.

Schmidt, R., Freidl, W., Fazekas, F., Reinhart, B., Grieshofer, P., Koch, M., et al. (1994). The Mattis Dementia Rating Scale: Normative data from 1,001 healthy volunteers. *Neurology*, *44*(5), 964–966.

Seidenberg, M., Hermann, B., Wyler, A. R., Davies, K., Dohan, F. C., & Laveroni, C. (1998). Neuropsychological outcome following anterior temporal lobectomy in patients with and without the syndrome of mesial temporal lobe epilepsy. *Neuropsychology*, *12*, 303–316.

Smith, A. (1982). *Symbol Digit Modalities Test manual (Rev. ed.)*. Los Angeles, CA: Western Psychological Services.

Stern, R. A., & White, T. (2001). *Neuropsychological Assessment Battery*. Lutz, FL: Psychological Assessment Resources.

Strauss, E., Sherman, M. S. E., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration norms, and commentary* (3rd ed.). New York: Oxford University Press.

Taylor, M. J., & Heaton, R. K. (2001). Sensitivity and specificity of the WAIS-III/WMS-III demographically corrected factor scores in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *7*, 867–874.

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*(4), 357–369.

The Psychological Corporation. (2002). *WAIS-III–WMS-III–WIAT-II scoring assistant (rev.)*. San Antonio, TX: The Psychological Corporation.

Tombaugh, T. N. (2005). Test–retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Archives of Clinical Neuropsychology*, *20*(4), 485–503.

Turf, E., Ingrisawang, L., Turf, M., Ball, J. D., Stutts, M., Taylor, J., et al. (1999). A cohort study to determine the epidemiology of estuary-associated syndrome. *Virginia Journal of Science, 50*, 299–310.

Vanderploeg, R. D., Schinka, J. A., Jones, T., Small, B. J., Graves, A. B., & Mortimer, J. A. (2000). Elderly norms for the Hopkins Verbal Learning Test–Revised. *Clinical Neuropsychologist*, *14*(3), 318–324.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scales – Third edition*. San Antonio, TX: The Psychological Corporation.

Wilkinson, G. S. (1993). *Wide Range Achievement Test 3 – Administration manual*. Wilmington, DE: Jastak Associates, Inc.