Published in final edited form as:
J Clin Exp Neuropsychol. 2011 June ; 33(5): 505–522. doi:10.1080/13803395.2010.535504.

Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change

Lucette A. Cysique[1,2], **Donald Franklin Jr**[1], **Ian Abramson**[1], **Ronald J. Ellis**[1], **Scott Letendre**[1], **Ann Collier**[3], **Christina Marra**[3], **David Clifford**[4], **Benjamin Gelman**[5], **Justin McArthur**[6], **Susan Morgello**[7], **David Simpson**[7], **J. Allen McCutchan**[1], **Igor Grant**[1], **Robert K. Heaton**[1], and **CHARTER group, and the HNRC group**

[1]HIV Neurobehavioral Research Center (HNRC; http://hnrc.hivresearch.ucsd.edu/), Department of Psychiatry, University of California at San Diego, San Diego, California

[2]Brain Sciences, School of Psychiatry, Faculty of Medicine, University of New South Wales, Sydney, Australia

[3]University of Washington, Seattle; Seattle, WA

[4]Washington University, St. Louis; St. Louis, MO

[5]University of Texas Medical Branch; Galveston, TX

[6]Johns Hopkins University; Baltimore, MD

[7]Mount Sinai School of Medicine; New York, NY

Abstract

Reliable detection and quantification of longitudinal cognitive change are of considerable importance in many neurological disorders, particularly to monitor central nervous system effects of disease progression and treatment. In the current study, we developed normative data for repeated neuropsychological (NP) assessments (6 testings) using a modified Standard Regression-Based (SRB) approach in a sample that includes both HIV-uninfected (HIV−, N=172) and neuromedically stable HIV-infected (HIV+, N=124) individuals. Prior analyzes indicated no differences in NP change between the infected and uninfected participants. The norms for change included correction for factors found to significantly affect follow-up performance, using hierarchical regression. The most robust and consistent predictors of follow-up performance were the prior performance on the same test (which contributed in all models) and a measure of prior overall NP competence (predictor in 97% of all models). Demographic variables were predictors in 10%-46% of all models and in small amounts; while test retest interval contributed in only 6% of all models. Based on the regression equations, standardized change scores (z-scores) were computed for each test measure at each interval; these z scores were then averaged to create a total battery change score. An independent sample of HIV− participants who had completed 8 of the 15 tests was used to validate an abridged summary change score. The normative data are available in an electronic format by email request to the first author. Correction for practice effects based on normative data improved the consistency of NP impairment classification in a clinically stable longitudinal cohort after baseline.

Correspondence: **Robert K. Heaton**: Department of Psychiatry, University of California San Diego, 9500 Gilman Drive #0603, La Jolla, California, 92093-0603. Phone 858-534-4044; Fax: 858-534-9917. rheaton@ucsd.edu.
Request of electronic normative data: lcysique@ucsd.edu and specify “request of change normative data” in the subject head of the email. The data will also be available from the HNRC website: http://hnrc.hivresearch.ucsd.edu/

Keywords

Normative data; longitudinal studies; regression; regression change score; SRB; practice effect

Introduction

The development of norms to reliably identify and quantify neuropsychological (NP) impairment has had a significant impact for clinical and research neuropsychology (Strauss, Sherman, & Spreen, 2006). Normative data that corrects for demographic factors associated with NP performance in healthy controls (i.e., age, education, gender and sometimes ethnicity or pre-morbid ability), enables accurate estimation of disease prevalence, clearer definition of disease and treatment effects and improved clinical management in a variety of neurological and psychiatric disorders.

Neuropsychologists also have emphasised the necessity for providing norms in order to reliably estimate cognitive change over time (Heaton, et al., 2001; Salthouse & Tucker-Drob, 2008). Identifying real cognitive change in individuals undergoing repeated NP assessment is potentially valuable for both clinical and research purposes, but diagnosing NP change beyond baseline assessments is difficult because of random biological variation, measurement errors, and practice effects (McCaffrey & Westervelt, 1995)

“Practice effect” (PE), or “learning effect” is seen on most NP measures and is larger on measures of problem solving or tests with a high novelty component (Dikmen, et al., 1999). PE complicates detection of meaningful change and leads to underdetection of impairment beyond baseline assessment even when alternate test forms are available.

Recent research advances in the area of repeated NP assessment have shown that prediction of cognitive change is substantially improved in clinical samples, including samples for which cognitive change is expected (after treatment for example), when referenced against normative standards. Normative standards are ideally derived from appropriate comparison samples for which no change is expected beyond normal test re-test fluctuation (including practice effect) or regression towards the mean (Heaton, et al., 2001; Salthouse & Tucker-Drob, 2008).

One method that has been extensively validated in healthy as well as clinical samples, and that is now commonly used in test development, is the Standard Regression-Based (SRB) change score approach (M. R. Basso, Carona, Lowery, & Axelrod, 2002; Brandt & Benedict, 2001; Chelune, Naugle, Lüders, Sedlak, & Awad, 1993; Dikmen, Heaton, Grant, & Temkin, 1999; Martin, et al., 2002; McSweeny, Naugle, Chelune, & Luders, 1993). This method has an advantage over other change score approaches (Collie, Darby, Falleti, Silbert, & Maruff, 2002) in that it can accommodate multivariate modelling (Temkin, Heaton, Grant, & Dikmen, 1999) with inclusion of numerous factors that may influence NP performance over several assessments.

Longitudinal norms for cognitive change using versions of the SRB change score approach have involved widely used NP measures in healthy and relatively young volunteers (Attix, et al., 2009; M. R. Basso, et al., 2002; Dikmen, et al., 1999; Levine, Miller, Becker, Selnes, & Cohen, 2004; Martin, et al., 2002; Temkin, et al., 1999), elderly samples (Duff, et al., 2008; Duff, et al., 2005; Frerichs & Tuokko, 2005; Ivnik, et al., 1999) and samples at risk for cognitive impairment (Heaton, et al., 2001; Hermann, et al., 1996; Sawrie, Chelune, Naugle, & Luders, 1996). Most include normative formulas for individual tests. A few studies have provided norms for scores representing cognitive domains (Duff, et al., 2005) or cognitive

factor scores derived from specific NP batteries (Ivnik, et al., 1999), but none of these studies provides guidance to detect and quantify overall cognitive change as derived from a total NP test battery involving multiple cognitive domains.

In this regard, the study of Woods, et al. (2006) proposed a battery change score approach using the Reliable Change Index method (RCI). As noted by the authors, a strong advantage of a summary change score approach is to improve the reliability and validity of the observed change as it is based on multiple cognitive abilities. In fact, cognitive change across several NP functions, and in variable patterns, is common in many neurological disorders.

In the present study we have modified the SRB change score approach to allow for: 1) selection of predictors of cognitive change based on previous literature, but with the addition of an estimate of overall prior NP competence and, 2) development of a *summary* change score approach on the total NP test battery. We developed norms in samples of HIV-uninfected (HIV−) people and HIV-infected (HIV+) individuals who were carefully and independently classified as stable with regard to both disease-related and treatment-related neuromedical status. We validated these norms in another sample of HIV− individuals who took a subset of the test battery two or three times. Finally, to enhance accuracy in classifying impairment (versus change) after baseline, we provide normative corrections for practice effects and demonstrate that their use greatly reduces a tendency for clinically stable examinees to appear to have improved abilities over time.

The first aim of the current study was to develop normative longitudinal data based on multiple cognitive domains and defined by a single summary change score. To reach this aim, we initially used an HIV-negative (HIV−) sample to develop longitudinal normative SRB equations. The SRB equations were then applied to the clinically stable HIV+ group. The application of the norms to the HIV+ group was done to demonstrate that clinically stable HIV+ do not differ in terms of performance change as compared to HIV− individuals despite a slightly higher rate of baseline impairment rate.

Because the HIV+ and HIV− cohorts showed comparable test-retest results, combining their data for normative purposes provides the advantages of a larger sample size and a broader range of baseline performances from individuals who also share general personal/background characteristics with the individuals to which the norms for change would be applied (e.g., HIV+ patients who are progressing off treatment, failing their current treatment or initiating new regimens for example). We believe this is the first study to develop norms for change in this context. The normative formulas will be made available in an electronic format, for ease of use. As noted, we also provide a method for correcting for PE on the six repeated assessments.

Methods

Participants

Of 296 volunteers at the San Diego HIV Neurobehavioral Research Center (HNRC) who completed the same test battery over two to six visits, 172 were HIV− controls recruited between 1999 and 2006 and 124 were clinically stable HIV+ individuals recruited through the multi-site CNS HIV Antiretroviral Therapy Effects Research (CHARTER) study between 2001 and 2007. An additional, “validation group” was 111 HIV− volunteers who completed a subset of the same test battery at two time points and 67 who completed the abbreviated battery at three time points at the HNRC between 1987 and 1999. Table 1 compares the demographic characteristics and provide summary results on the NP test battery for the normative HIV− and HIV+ groups and the validation sample.

Neuromedical stability was defined for the HIV+ group as having: (a) stable HIV disease indicators between visits (i.e., CD4 counts not changing among following categories >500, 500–200, <200; < 1 $\log_{10}$ change in their plasma HIV RNA levels (viral load); and no new AIDS defining illnesses), (b) no change in their antiretroviral regimen, and (c) no incident psychiatric illness (i.e., major depressive episode or substance use disorder) or neurological events (i.e., head injury or meningitis) between visits.

Participants with a history of non-HIV related neuromedical factors that might potentially cause neurocognitive impairment were excluded. These exclusion criteria were (a) head injury with unconsciousness greater than 30 minutes, (b) any known, non HIV related neurological disorders (e.g., epilepsy, stroke), psychotic disorders (schizophrenia) and (c) significant levels of current self-reported substance use, defined as more than three alcoholic drinks per day over the past 30 days, or use of any illegal drugs in the past 30 days. Diagnosis of bipolar disorder was not disqualifying so long as patients were stable on their medications.

Procedure

Neuropsychological Assessments—In the reference sample the NP test battery was composed of 15 individual NP tests which assessed 7 ability areas (see Table 2). When available, alternate versions of the NP tests (i.e., Hopkins Verbal Learning Test-Revised, Brief Visual Memory Test-Revised) were used according to the schedule detailed in the manual to minimize PE.

Data Analysis

Development of NP norms for change—To identify individuals who presented with overall NP change versus stability (“norms for change”), we used a statistical methodology based on the multivariate Standard Regression-Based (SRB) approach ((McSweeny, et al., 1993; Temkin, et al., 1999; see also Collie, et al., 2002, for review of change score methodologies). The advantage of multiple regression SRB is that it accounts for practice effect, regression towards the mean and other factors that potentially may influence test-retest variability in neurologically stable people (e.g., test-retest interval, demographics and “overall NP competence” at baseline) (Temkin, et al., 1999).

Prior to analysis, raw scores for each individual NP measure were transformed into scaled scores (with a mean of 10 and a standard deviation of 3 based on large samples (Heaton et al., 2004; Norman et al, submitted). The scaled score transformations provide a robust method to: a) normalize distributions of multiple NP tests using a common metric, b) attenuate the effects of outliers (Tabachnik & Fidell, 2007). The scaled score transformations will be part of the electronic file that can be requested by email to the first author.

The first step in developing norms for change consisted of statistically evaluating the comparability of the HIV+ and HIV− groups to determine whether they could be combined to create a larger *reference* group for which no NP change is expected beyond practice effect. The 172 HIV-controls were assessed at approximately one-year intervals (14.26 ± 4.89 months) and 124 neuromedically stable HIV+ individuals had been assessed at approximately six month intervals (7.28 ± 2.56 months). Baseline rate of impairment was 22% in the HIV− group and 28% in the HIV+ group (see Table 1) as defined by the Global Deficit Score (GDS) method (Carey et al., 2004).

Regression formulas were developed for the HIV− group *only*, using the method detailed in the following sections for the combined normative data. The change rate in the HIV+

individuals fell well between the 90% prediction intervals of stability based upon regression based norms developed with the 172 HIV-individuals as a reference (see Table 3). In no instance did the resulting rate of change (or no change) of the stable HIV+ group differ significantly from those of the HIV− control sample (all *p-values* >.20). Consistent with findings from previously published studies, test-retest intervals were not significantly associated with practice effect across testing visits for the current reference samples (Cysique, et al., 2009). Baseline impairment rate in the combined HIV+ and HIV− reference sample reached 24.6%. Table 3 presents the percent rate of change data on the mean scaled score across the study time points for the HIV+ and HIV− samples.

Additionally, we tested whether HIV status was associated with differences in performance change using a multivariate approach, which also provides an adjustment for attrition. For this we developed a mixed effect regression model. HIV status and time were included as fixed linear effects, HIV status and time as a fixed non-linear interaction effect, and subject as a random effect. By incorporating subject-specific effects, i.e. the random intercepts and slopes, the mixed-effects model has the advantage of removing biases in estimation due to attrition at later time points and by adjusting the groups, time and interaction effect intercepts and slopes of the model. We included the model described above (model 1) and a model with demographic and test-retest effect (model 2) in the appendix. In all instances, the HIV status and time interactions were not significant, indicating that there was no difference between the HIV− individuals and HIV+ individuals' performance changes over time.

The entire reference group was therefore composed of 296 volunteers at baseline and at time 2 (test-retest interval: 11.3 months ± 5.3); 241 at time 3 (test-retest interval from baseline: 21.8 months ± 9.0), 171 at time 4 (test-retest interval from baseline: 32.5 months ± 11.0), 132 at time 5 (test-retest interval from baseline: 42.3 months ± 13.2) and 64 at time 6 (test-retest interval from baseline: 49.8 months ± 16.5). To investigate differences in performance among subgroups across time, we conducted a series of repeated measure ANOVAs testing for (a) group effect (this effect was represented by including sub-samples according to how many sessions they completed), (b) time effect, and (c) group -by- time interaction effect. The interaction could indicate, for example, that the sample that had completed 6 sessions was different from those who had completed only 2 sessions. In all analyses the interaction term was non-significant (p>.90), indicating that NP performance was homogenous regardless of number of visits completed.

Table 4 presents the baseline scaled scores on the 15 NP measures for the reference sample and on the eight NP measures for the validation sample.

To determine which factors influenced longitudinal NP performance in the reference sample, we conducted a series of hierarchical regression analyses using the follow-up scaled score as the outcome variable and a predetermined list of candidate predictors. The selection of these factors and their order as hierarchical predictors of follow-up NP performance was based on previous research literature in the study of longitudinal NP performance (M. Basso, Bornstein, & Lang, 1999; Duff, et al., 2007; Heaton, et al., 2001; Levine, et al., 2004). This step is different from the original method (Chelune, et al., 1993) for which stepwise regression was used. The hierarchical method used the knowledge of previous findings to inform the order of entry into the regression, while stepwise uses only a sample specific statistical cut-off.

The predictor variables that were considered in the multivariate analyses included (in order): 1) baseline or previous test performance of the individual NP measure in question (scaled scores), 2) Baseline or previous NP competence (defined as the previous mean scaled score on the battery, with exclusion of the NP measure currently tested in the regression model);

3) age (in years), 4) education (in years); 5) gender (male vs. female coded as 0 and 1), 6) ethnicity (Caucasians vs. Others; coded as 1 and 0), and 7) relevant test-retest interval (in months). Any predictive factor that accounted for significant R^2 change above and beyond the previously entered predictors ($p<.10$) was retained in order to compute reference regression formulas. Table 5 lists which factors were found to be significant predictors of follow-up performance for all the 15 NP measures at all study time points. Multi-colinearity was not detected in any of the models when tolerance statistics and the magnitude of change in the overall R^2 were examined for each model after a particular factor was added or excluded.

Once the initial predictive models were built, we tested some interaction terms (i.e., overall competence * age, overall competence*previous test score; overall competence * education level, overall competence * sex and age*edu) chosen to detect possible interactions of demographic factors with overall NP competence. From all the models containing the interaction terms (as described above), at most one interaction of predictors emerged as marginally informative for the following NP tests: Letter Fluency and WCST perseverative errors. Accordingly, we decided not to encumber the models by retaining these few interactions.

Next, the significant predictors were retained in a series of standard multivariate regression analyses to derive regression formulas for each individual NP test. These formulas served to compute the 15 individual predicted follow-up scaled scores as shown below [formula 1]:

$$Y_p = \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n + a \quad [1]$$

In this formula Y_p is the predicted scaled score at *follow-up*, β_1 is the regression coefficient (slope) for predictor X_1; β_2 for predictor X_2; β_n for predictor X_n; X_1 is the observed baseline or previous test score, X_2 and X_n are the overall NP competence, and demographic or re-test interval factors entered into the model and a is the intercept.

To inspect for any heteroscedasticity of demographic effects in the resulting 15 models at the six time points (Tabachnick & Fidell, 2007), we conducted a series of Pearson correlation analyses between the residuals of each model and age, education, and gender separately. None of the analyses reached statistical significance ($p<.05$), showing that demographic effects were quite similar across the NP performance range and over time.

Finally, because most neuropsychological assessments involve the evaluation of several cognitive abilities often yielding a number of individual NP measures, we have developed a method that takes into account all the information within the test battery by computing a battery NP change score or a *summary NP change score*.

Summary regression-based change scores (sRCS) were computed as follows: The standard deviation of the residuals (i.e., error term of the regression model in SD units) for each of the 15 final regression models was computed for time 1 predicting time 2; time 2 predicting time 3; time 3 predicting time 4, time 4 predicting time 5; time 1 predicting time 3 and also predicting time 4, and time 5. Formulas for time 4 predicting time 5 and time 1 predicting time 5 were applied at time 6. This was applied as such because of the relatively small size of the sample at time 6. Altogether, 15 individual z-scores at 6 different time ranges were computed, by dividing the difference between predicted and obtained follow-up scaled scores by the error term of the regression model [see 2]. The resulting Z-score reflects how well or poorly the participant did at follow-up, relative to expectations for a neurocognitively stable person with his/her baseline (or previous) score and other variable-specific baseline predictors:

$$Z-score=(X_o - X_p)/(SD_{residual}) \quad [2]$$

In this formula, X_o is the scaled score at follow-up; X_p is the predicted follow-up scaled score derived from the reference sample regression equations [1], and $SD_{residual}$ is the standard deviation of the residuals from the reference sample regression model. Note that this regression based change score can be negative (if the obtained follow-up score is worse than the predicted) or positive (if the obtained score is greater than the predicted score).

The final step in the development of the norms for change was to average the 15 individual regression based change scores (z-scores), at relevant times, to compute the sRCS and determine a 90% confidence interval to define "no change" on a summary of the entire test battery. Thus, participants in the top 5% of the sRCS distribution of the reference sample were defined as "improved" and the bottom 5% were defined the "decliners". Cut-offs for significant change are also provided in the appendix for 80% and 70% confidence intervals.

Computation of the Practice effect (PE)—PE corrections were developed *separately from the norms for change* using scaled scores, on each individual NP measure across the study time points. The *median PE* of the normative sample was selected for this purpose. That is, the median PE (in scaled score units) was subtracted from the scaled score at follow-up, to estimate what the performance would have been without practice. We then applied these corrections to the mean scaled score to illustrate how it helps avoid spurious findings of apparent improvement (and decreased sensitivity to impairment or decline) in functioning over repeated NP assessments.

Missing data—Of the 18,000 possible scores on every NP test measure across the six testings, 135 (0.75%) were missing. However for each individual, the summary NP performance was composed of at least 12 individual NP measures, in our judgement providing a reliable estimate of global NP performance.

Validation—The norms for change (i.e., prediction formulas from the reference sample regression models and resulting z-scores and sRCS at appropriate times) were then applied to a sample the 111 HIV-persons for validation. These norms were applied to an abridged NP battery consisting of eight of the initial 15 NP tests which were available for the validation sample (Table 2 and Table 3). For this a separate sRCS and confidence interval were computed on the abridged NP battery, using formulas derived from the reference group for the eight relevant NP tests. Test-retest interval was not involved in these analyses because it had no predictive power in the hierarchical models in the reference sample on any of the eight NP measures concerned.

Analyses were conducted using PAWS 18.0 version, JMP 7.0 version (SAS Inc) and the effect size calculator from (Lipsey & Wilson, 2001).

Results

Reliability and practice effect (PE) findings

Table 6 presents median overall PE computed from scaled scores (i.e., using the same metric across all NP measures). We retained the "median PE" rather than the alternative "average PE" because it provided a better estimate of overall stability. Indeed, the average PE across all 15 measures tended to overestimate PE in the mean scaled score (the most reliable measure; see Table 7) across time, while the median PE correction better matched the goal of "no change".

In Figure 1, we illustrate the application of these PE corrections on the mean scaled score for the sample re-tested five times after baseline (see also Table 7 which provides this for all time points). Without correction, an apparent improvement in the mean scaled score is observable at each session reaching almost a full scaled score point at time 6. PEs were found to be larger, in most instances, at the first test-retest interval (sum corrections = 6 scaled score points), but considerable additional PE also occurred from time 2 to time 5 (sum corrections = 6.5 additional scaled score points, see Table 5 and Figure 1). The median PE figures did not provide any additional correction between time 5 and time 6 (see sum in Table 5 suggesting that PE is subtle or variable thereafter).

Table 8 presents reliability statistics in the reference sample. Test re-test reliability was overall adequate. Median correlation coefficients (r's) across the 15-test battery were .67 for time 1 – 2; .68 for time 2 – 3; .70 for time 3 – 4; .69 for time 4 – 5, and .68 for time 5 – 6. The highest test retest reliability was observed for the mean scaled scores (.88 – .92), and for certain speed of information processing tests and working memory tests (Digit Symbol: .89; PASAT-50: .82). The reliability was lowest for the delayed recall in memory tests (medians of .52 and .63). The use of scaled scores rather than raw scores improved the reliability statistic on several tests, while in other tests the difference with the raw scores was small in either direction (data not shown).

Validation of the sRCS

In the validation sample, classification of cases as improving or declining at time 2 or 3 did not differ from the 5% reference sample's prediction (Table 9). When considering the continuous sRCS, the validation sample declined more than the reference sample only at time 2. While this represented a significant statistical difference, we determined that this average decline (−0.16) was far from the confidence interval boundary for significant decline, that is > 0.825. Thus, use of confidence intervals to determine a significant and clinically meaningful level of change at the individual level remains quite accurate; that is, the "percent declined" in this validation sample at time 2 is only 6% (very close to the 5% in the reference sample) and the maximum deviation from expected across these validation points was two percent (Table 8).

Further analyses using each of the eight individual NP measures yielded somewhat different proportions of stability vs. change (improvement or decline) especially on the Trail Making Test part B. Thus, targeting the *summary* measure of cognition for interpretation of meaningful change (here the sRCS) approach, while most conservative, is also more reliable and should be recommended over using the SRB approach on individual tests.

Finally, we compared the continuous and the discrete predictions of change in the reference sample using the 15 NP test-based sRCS versus the 8 NP test-based sRCS. We found for time 1 – 2 a 90.5% agreement in the classification of change and a significant correlation for the continuous sRCS ($r = .77$; $p < .0001$); for time 2 – time 3 there was an 88.0% agreement in the classification of change and a significant correlation for the continuous sRCS ($r = .79$; $p < .0001$); and for time 1 – 3 there was a 91.0% agreement in the classification of change and a significant correlation for the continuous sRCS ($r = .81$; $p < .0001$). In all instances, disagreement in change classification involved "change" versus "no change" (never change in the opposite direction).

Discussion

In the current study we developed a series of summary regression-based change scores (sRCS) over six NP testings in a sample of 296 HIV-uninfected and neuromedically stable HIV-infected individuals. These longitudinal normative standards - in an abridged format -

were then successfully applied to validation samples of 111 HIV-uninfected individuals who were assessed twice or 3 times (N=67) based on 8 NP tests. Estimates of the proportion of individuals with significant NP performance change did not differ between the normative sample and the validation sample, in both groups approximately 5% improved or declined and 90% were stable. We also found that the 15 NP test-based norms and the 8 test-based norms were highly correlated demonstrating a high level of agreement in the classification of “change” or “no change”.

Prior to the development of our norms we have demonstrated, using two different statistical methods (the SRB method and mixed effect regression analyses) that clinically stable HIV+ individuals perform similarly to HIV− individuals in terms of test-retest change despite a slightly higher impairment rate at baseline in the HIV+ sample. Our study confirms and extends our previous findings (Cysique et al., 2009) where we had found that clinically stable HIV+ individuals performed similarly to HIV− individuals over a one-year period. Here we show that this longitudinal comparability extended to a three-year period, and was even seen when we adjust for attrition as allowed mixed effect regression analyses. Another large study has found that clinically stable HIV+ individuals had stable psychomotor speed performance over a five-year period as compared to HIV− individuals (Cole et al., 2007). There is therefore cumulative evidence that the stability of performance across several years is associated with stable HIV disease, providing a valuable pool of performance data that can be standardised use to create norms for change. While this applies to HIV infection, where long-term treatment is known to be increasingly effective and increasing patients’ life expectancies almost as long as non-HIV infected individuals, we would recommend as proposed here, that the stability of performance is empirically determined (by comparison to HIV− controls) prior to normative data development.

Moreover it should be noted that the HIV+ reference sample represents a carefully selected subgroup of the CHARTER cohort, who have both minimal confounds and very stable disease and treatment over time as well as no incidence of psychiatric conditions and substance use disorders. Because of this, their baseline impairment rate is much lower than the rest of CHARTER cohort (28% versus 52% overall and 40% in the total subgroup with minimal comorbidity (Heaton et al. press). This explains why we did not find a statistically significant difference (at $p<.05$) in baseline impairment and overall NP performance on the GDS when compared to the HIV− reference sample (only a trend; $p<.08$). Concerning the HIV− reference sample, it should be noted that in the current era, HIV− participants in HNRC studies (those who have the current NP battery) share lifestyle factors with the HIV+ reference sample (e.g., they do not have current or incident substance use disorders but could have lifetime disorders). Our combined HIV+ and HIV− sample provides larger age and educational ranges, as well as broader gender and ethnicity representations (see Table 1) and increase range of baseline NP performance. All those characteristics are not only valuable for HIV research, but also beyond this field of research, as the methods should be applicable to a range of neuromedical conditions. Further inspection of our normative data demonstrated that our norms for change were independent of baseline impairment. Indeed, the sRCS were at all times not statistically different between impaired and unimpaired individuals at baseline (dichotomous sRCS $p>.28$). When using the continuous sRCS, there was no systematic indication that the impaired individuals declined or improved as a group as compared to the unimpaired group.

In addition to this evidence of validity of the current approach in clinically stable HIV+ patients and HIV-controls, the same methods were found to be valid for detecting cognitive changes produced by HIV-associated disease and a CNS penetration of antiretrovirals in another study (Cysique, et al., 2009). We have also used the same sRCS methodology in

cross-cultural settings, finding that cognitive decline measured in this way was sensitive to HIV disease-related change in China (Cysique, et al., 2010).

Another method for detecting meaningful cognitive change is based on PRESS statistics (Berres, Zehnder, Blasi, & Monsch, 2008.; Blasi, et al., 2009). Briefly, this method reportedly provides an improved selection of prediction models compared to the stepwise regression often used prior the generation of SRB change scores. Interestingly, several of their models include interaction terms involving demographics. In our study, we found that the inclusion of interaction terms did not substantially improve our predictive models. Some important differences between the methods in the two studies include that the PRESS-based method was validated on only one test of learning and memory (i.e., the California Verbal Learning Test) and did not account for overall NP competence as we did (of course, our measure of prior NP competence requires longitudinal use of a test battery, rather than a single test). Without the inclusion of overall NP competence, a significant amount of variance may be unaccounted in predicting NP change. This may explain why the addition of interaction terms is important in the PRESS studies, but not in ours. In future studies use of the *exact* same predictive models (models inclusive of overall NP competence as a predictor when using battery wide summary change scores) might be compared for the SRB method and the PRESS method.

With respect to our inclusion of overall NP competence, one might assume that a pre-morbid ability estimate (e.g. pre-morbid IQ) might perform better than current cognitive functioning. However, in one prior study, it has been shown that a measure of premorbid functioning (estimated IQ) did not independently predict meaningful cognitive change (Frerichs & Tuokko, 2005).

It might be suggested that the construction of our SRB formula privileged the prior performance in a test and the overall NP competence as they were entered as first and second predictors in the hierarchical model which eventually served to select the predictors to be included in the final formulas. In fact, we found that overall NP competence was a significant predictor in 97% of all models, while demographic factors were less frequently significant (see Table 5; age is a predictor in 46% of all models; education is a predictor in 15% of all models, gender is a predictor in 13% of all models; ethnicity is a predictor in 10% of all models; and test-rested interval is a predictor in only 6% of all models). In addition, the variance explained by both the prior performance on the same test and the overall NP competence contributed to 70–90% of the overall model R^2; while demographic factors accounted for between 5–15% of explained variance, and test retest-interval to 1.5–3%. Further comparison of our “order of entry” of demographics confirmed that the variance explained by demographics was lessened in our approach, but that our final models were simpler and less prone to apparently spurious associations (e.g., predictions going in inconsistent directions across the test variables; see Table A2 in the appendix).

While overall test-retest reliability was adequate for individual tests, the highest reliability was found for the mean scaled scores and some psychomotor speed and working memory tests. Lowest reliability was found for some delayed recall and executive functions measures. These patterns are similar to findings of previous studies (M. Basso, et al., 1999; Temkin, et al., 1999). Higher reliability in psychomotor speed based tests is known to be boosted by their psychometric properties such as large range of possible values and approximation to a normal distribution even on raw scores. In contrast, measures of memory and executive functions often have a restricted range of possible values, especially in follow-up assessments. They are also intrinsically more susceptible to learning effects (M. Basso, et al., 1999; Temkin, et al., 1999). We found the improved distributions provided by the scaled scores enhanced the test-retest reliability of those measures and did not

substantially change those of other NP measures. In the case of the memory tests, alternate versions were used for each of the five follow-ups, perhaps contributing to reduced test-retest reliability.

Practice effects were found to be larger, in most instances, in the first test-retest interval in accordance with previous observations (Collie, et al., 2003). However, additional PEs were found at visits three to five for multiple test measures and in particular on the WCST, demonstrating that learning continues over multiple exposures to tests of executive functioning (M. Basso, et al., 1999; Levine, et al., 2004).

Our study had several limitations:

1) The reference group was mainly male, but with a relatively good age range (18–66) and education range (7–20), and was diverse in terms of ethnicity representation (Table 1). Although age and gender were not strong predictors of test-battery test change in our models, our results maybe less generalizable to women and older individuals (beyond age 65).

2) The sRCS method is relatively complex, requiring several computational steps. To aid potential users, we have built a password protected spreadsheet which will include the needed raw to scaled scores transformations, the computation of the z-scores and the sRCS with the normative confidence intervals. Only the final computed sRCS will appear on this spreadsheet and the entry field of the relevant predictor for each of the 15 NP measures. The formulas will be password protected and accessible in plain text format upon request to the first author. The spreadsheet is available upon request to the first author (lcysique@ucsd.edu) and will be held by the HIV Neurobehavioral Research Center at San Diego, USA < http://hnrc.hivresearch.ucsd.edu/>.

3) The NP testing required is relatively extensive, requiring approximately 2 hours to complete. The optimal use of the current sRCS would require the use of the 15 NP measures in the version presented in Table 3. However, the abridged sRCS based on 8 NP measures might be used as a second choice as we have shown that it has substantial overlap with the 15 NP measure-based sRCS predictions.

In conclusion, we provide norms for a moderately comprehensive test battery, using an improved method to quantify global cognitive change that can be applied in clinical samples to determine disease-related incidence of cognitive decline or incidence of cognitive improvement after treatment initiation. In the future, the summary sRCS approach should be compared to other recently developed methods for interpretation of cognitive change and to determine how well they perform overall and within specific cognitive domains.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the HIV Neurobehavioral Research Center (HNRC* supported by Center award MH 62512 from NIMH), the CNS HIV Anti-Retroviral Therapy Effects Research (CHARTER**) study (supported by award N01 MH22005 from the National Institutes of Health), by the MH 58076 grant, and Brain Sciences post-doctoral fellowship at the University of New South Wales.

Dr. Cysique is supported by the Brain Sciences post-doctoral fellowship from the University of New South Wales, Sydney, Australia.

References

Attix DK, Story TJ, Chelune GJ, Ball JD, Stutts ML, Hart RP, et al. The prediction of change: normative neuropsychological trajectories. The Clinical Neuropsychologist. 2009; 23(1):21–38. [PubMed: 18720272]

Basso M, Bornstein R, Lang J. Practice effects on commonly used measures of executive function across twelve months. The Clinical Neuropsychologist. 1999; 13(3):283–292. [PubMed: 10726600]

Basso MR, Carona FD, Lowery N, Axelrod BN. Practice effects on the WAIS-III across 3- and 6-month intervals. The Clinical Neuropsychologist. 2002; 16(1):57–63. [PubMed: 11992227]

Berres M, Zehnder A, Blasi S, Monsch AU. Evaluation of diagnostic scores with adjustment for covariates. Stat Med. 2008; 27(10):1777–1790. [PubMed: 17968872]

Blasi S, Zehnder AE, Berres M, Taylor KI, Spiegel R, Monsch AU. Norms for change in episodic memory as a prerequisite for the diagnosis of mild cognitive impairment (MCI). Neuropsychology. 2009; 23(2):189–200. [PubMed: 19254092]

Brandt, J.; Benedict, RHB. Hopkins Verbal Learning Test-Revised. Lutz, FL: Psychological Assessment Resources, Inc; 2001.

Carey CL, Woods SP, Gonzalez R, Conover E, Marcotte TD, Grant I, et al. Predictive validity of global deficit scores in detecting neuropsychological impairment in HIV infection. Journal of Clinical and Experimental Neuropsychology. 2004; 26(3):307–319. [PubMed: 15512922]

Chelune GJ, Naugle RI, Lüders H, Sedlak J, Awad IA. Individual change following epilepsy surgery: Practice effects and base-rate information. Neuropsychology. 1993; 1:41–52.

Cole MA, Margolick JB, Cox C, Li X, Selnes OA, Martin EM, et al. Longitudinally preserved psychomotor performance in long-term asymptomatic HIV-infected individuals. Neurology. 2007; 69(24):2213–2220. [PubMed: 17914066]

---

*The San Diego HIV Neurobehavioral Research Center [HNRC] group is affiliated with the University of California, San Diego, the Naval Hospital, San Diego, and the Veterans Affairs San Diego Healthcare System, and includes: Director: Igor Grant, M.D.; Co-Directors: J. Hampton Atkinson, M.D., Ronald J. Ellis, M.D., Ph.D., and J. Allen McCutchan, M.D.; Center Manager: Thomas D. Marcotte, Ph.D.; Naval Hospital San Diego: Braden R. Hale, M.D., M.P.H. (P.I.); Neuromedical Component: Ronald J. Ellis, M.D., Ph.D. (P.I.), J. Allen McCutchan, M.D., Scott Letendre, M.D., Edmund Capparelli, Pharm..D., Rachel Schrier, Ph.D.; Neurobehavioral Component: Robert K. Heaton, Ph.D. (P.I.), Mariana Cherner, Ph.D., Steven Paul Woods, Psy.D.; Neuroimaging Component: Terry Jernigan, Ph.D. (P.I.), Christine Fennema-Notestine, Ph.D., Sarah L., Archibald, M.A., John Hesselink, M.D., Jacopo Annese, Ph.D., Michael J. Taylor, Ph.D., Neurobiology Component: Eliezer Masliah, M.D. (P.I.), Ian Everall, FRCPsych., FRCPath., Ph.D.,Cristian Achim, M.D., Ph.D..; Neurovirology Component: Douglas Richman, M.D., (P.I.), David M. Smith, M.D.; International Component: J. Allen McCutchan, M.D., (P.I.); Developmental Component: Ian Everall, FRCPsych., FRCPath., Ph.D (P.I.), Stuart Lipton, M.D., Ph.D.; Clinical Trials Component: J. Allen McCutchan, M.D., J. Hampton Atkinson, M.D., Ronald J. Ellis, M.D., Ph.D., Scott Letendre, M.D.; Participant Accrual and Retention Unit: J. Hampton Atkinson, M.D. (P.I.), Rodney von Jaeger, M.P.H.; Data Management Unit: Anthony C. Gamst, Ph.D. (P.I.), Clint Cushman, B.A., (Data Systems Manager), Daniel R. Masys, M.D. (Senior Consultant); Statistics Unit: Ian Abramson, Ph.D. (P.I.), Christopher Ake, Ph.D., Florin Vaida, Ph.D.
The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

**The CNS HIV Anti-Retroviral Therapy Effects Research (CHARTER) group is affiliated with the Johns Hopkins University, Mount Sinai School of Medicine, University of California, San Diego, University of Texas, Galveston, University of Washington, Seattle, Washington University, St. Louis and is headquartered at the University of California, San Diego and includes: Director: Igor Grant, M.D.; Co-Directors: J. Allen McCutchan, M.D., Ronald J. Ellis, M.D., Ph.D., Thomas D. Marcotte, Ph.D.; Center Manager: Donald Franklin, Jr.; Neuromedical Component: Ronald J. Ellis, M.D., Ph.D. (P.I.), J. Allen McCutchan, M.D., Terry Alexander, R.N.; Laboratory, Pharmacology and Immunology Component: Scott Letendre, M.D. (P.I.), Edmund Capparelli, Pharm.D.; Neurobehavioral Component: Robert K. Heaton, Ph.D. (P.I.), J. Hampton Atkinson, M.D., Steven Paul Woods, Psy.D., Matthew Dawson; Virology Component: Joseph K. Wong, M.D. (P.I.); Imaging Component: Christine Fennema-Notestine, Ph.D. (P.I.), Terry L., Jernigan, Ph.D., Michael J. Taylor, Ph.D., Rebecca Theilmann, Ph.D.; Data Management Unit: Anthony C. Gamst, Ph.D. (P.I.), Clint Cushman,; Statistics Unit: Ian Abramson, Ph.D. (P.I.), Florin Vaida, Ph.D.; Protocol Coordinating Component: Thomas D. Marcotte, Ph.D. (P.I.), Rodney von Jaeger, M.P.H.; Johns Hopkins University Site: Justin McArthur (P.I.), Mary Smith; Mount Sinai School of Medicine Site: Susan Morgello, M.D. (Co-P.I.) and David Simpson, M.D. (Co-P.I.), Letty Mintz, N.P.; University of California, San Diego Site: J. Allen McCutchan, M.D. (P.I.), Will Toperoff, N.P..; University of Washington, Seattle Site: Ann Collier, M.D. (Co-P.I.) and Christina Marra, M.D. (Co-P.I.), Trudy Jones, M.N., A.R.N.P.; University of Texas, Galveston Site: Benjamin Gelman, M.D., Ph.D. (P.I.), Eleanor Head, R.N., B.S.N.; and Washington University, St. Louis Site: David Clifford, M.D. (P.I.), Muhammad Al-Lozi, M.D., Mengesha Teshome, M.D.
The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Government.

Collie A, Darby DG, Falleti MG, Silbert BS, Maruff P. Determining the extent of cognitive change after coronary surgery: a review of statistical procedure. Annals of Thoracic Surgery. 2002; 73:2005–2011. [PubMed: 12078822]

Collie A, Maruff P, Darby D, McStephen M. 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. 2003; 9(3):419–428. [PubMed: 12666766]

Cysique LA, Letendre SL, Ake C, Jin H, Franklin DR, Gupta S, et al. Incidence and nature of cognitive decline over 1 year among HIV-infected former plasma donors in China. AIDS. 2010; 24(7):983–990. [PubMed: 20299964]

Cysique LA, Vaida F, Letendre S, Gibson S, Cherner M, Woods SP, et al. Dynamics of cognitive change in impaired HIV-positive patients initiating antiretroviral therapy. Neurology. 2009; 73(5): 342–348. [PubMed: 19474412]

Dikmen SS, Heaton RK, Grant I, Temkin NR. Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. Journal of the International Neuropsychological Society. 1999; 5(4):346–356. [PubMed: 10349297]

Duff K, Beglinger LJ, Schultz SK, Moser DJ, McCaffrey RJ, Haase RF, et al. Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. Archives of Clinical Neuropsychology. 2007; 22(1):15–24. [PubMed: 17142007]

Duff K, Beglinger LJ, Van Der Heiden S, Moser DJ, Arndt S, Schultz SK, et al. Short-term practice effects in amnestic mild cognitive impairment: implications for diagnosis and treatment. International Psychogeriatrics. 2008; 20(5):986–999. [PubMed: 18405398]

Duff K, Schoenberg MR, Patton D, Paulsen JS, Bayless JD, Mold J, et al. Regression-based formulas for predicting change in RBANS subtests with older adults. Archives of Clinical Neuropsychology. 2005; 20(3):281–290. [PubMed: 15797165]

Frerichs RJ, Tuokko HA. A comparison of methods for measuring cognitive change in older adults. Archives of Clinical Neuropsychology. 2005; 20(3):321–333. [PubMed: 15797168]

Heaton R, Temkin N, Dikmen S, Avitable N, Taylor M, Marcotte T, et al. Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. Archives of Clinical Neuropsychology. 2001; 16(1):75–91. [PubMed: 14590193]

Heaton, RK.; Miller, SW.; Taylor, MJ.; Grant, I. Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults Scoring Program. Odessa: FL: Psychological Assessment Resources; 2004.

Heaton RK, Clifford D, Franklin DR, Woods SP, Ake C, Vaida F, et al. HIV-associated neurocognitive disorders persist in the era of potent antiretroviral therapy: CHARTER Study. Neurology. (In press).

Hermann BP, Seidenberg M, Schoenfeld J, Peterson J, Leveroni C, et al. Empirical techniques for determining the reliability, magnitude, and pattern of neuropsychological change after epilepsy surgery. Epilepsia. 1996; 37(10):942–950. [PubMed: 8822692]

Ivnik RJ, Smith GE, Lucas JA, Petersen RC, Boeve BF, Kokmen E, et al. Testing normal older people three or four times at 1- to 2-year intervals: Defining normal variance. Neuropsychology. 1999; 13(1):121–127. [PubMed: 10067783]

Levine A, Miller E, Becker J, Selnes O, Cohen BA. Normative data for determining significance of test-retest differences on eight common neuropsychological instruments. The Clinical Neuropsychologist. 2004; 18:373–384. [PubMed: 15739809]

Lipsey, M.; Wilson, D. Practical meta-analysis. London: Sage Publications; 2001.

Martin R, Sawrie S, Gilliam F, Mackey M, Faught E, Knowlton R, et al. 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. 2002; 43(12):1551–1558. [PubMed: 12460258]

McCaffrey R, Westervelt H. Issues associated with repeated neuropsychological assessment. Neuropsychology Review. 1995; 5(3):203–221. [PubMed: 8653109]

McSweeny AJ, Naugle RI, Chelune GJ, Luders H. “T scores for change”: An illustration of a regression approach to depicting change in clinical neuropsychology. The Clinical Neuropsychologist. 1993; 7:300–312.

Norman MA, Moore DJ, Taylor M, Franklin D, Cysique LA, Ake C, et al. Demographically Corrected Norms for African Americans and Caucasians on the Hopkins Verbal Learning Test-Revised, Brief Visuospatial Memory Test-Revised, Stroop Color and Word Test, and Wisconsin Card Sorting Test 64-Card Version. Journal of Clinical and Experimental Neuropsychology. (Submitted).

Salthouse TA, Tucker-Drob EM. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology. 2008; 22(6):800–811. [PubMed: 18999354]

Sawrie SM, Chelune GJ, Naugle RI, Luders HO. Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. Journal of the International Neuropsychological Society. 1996; 2(6):556–564. [PubMed: 9375160]

Strauss, E.; Sherman, EMS.; Spreen, O. A Compendium Of Neuropsychological Tests: Administration, Norms, And Commentary. 3rd ed.. Oxford: Oxford University Press; 2006.

Tabachnik, BG.; Fidell, LS. Using Multivariate Statistics. 5th ed.. Boston: Person International Edition; 2007.

Temkin N, Heaton R, Grant I, Dikmen S. Detecting significant change in neuropsychological test performance: a comparison of four models. Journal of the International Neuropsychological Society. 1999; 5(4):357–369. [PubMed: 10349298]

Woods SP, Childers M, Ellis RJ, Guaman S, Grant I, Heaton RK. A battery approach for measuring neuropsychological change. Archives of Clinical Neuropsychology. 2006; 21(1):83–89. [PubMed: 16169705]

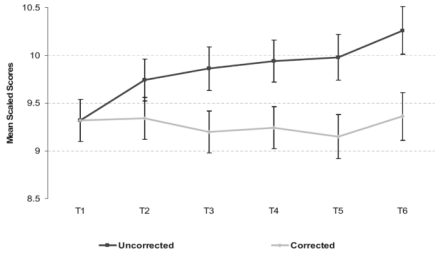

Figure 1.

Table 1

Demographic and clinical characteristics of the study samples

	HIV− Reference Sample (N=172)	HIV+ Reference Sample (N=124)	Total Reference Sample (HIV− & HIV+) (N=296)	HIV− Validation Sample (N=111)	*P*[1]	*P*[2]
Age	37.86 (11.30)	43.93 (8.89)	40.40 (10.8)	33.95 (8.12)	<.0001	<.0001
Education	12.92 (2.35)	12.85 (2.27)	12.9 (2.32)	14.30 (2.57)	.80	<.0001
Gender (% male)	64.5%	78.2%	70%	77%	.01	.19
Ethnicity (% Caucasian)	68%	44%	58%	66%	<.0001	.19
% HIV+ clinically stable	-	100%	41.8%	0%	-	-
% Baseline NP impairment	22%	28%	25%	15%	.22	.05
Overall impairment (GDS)	0.32 (0.38)	0.41 (0.48)	0.36 (0.43)	0.29 (0.32)	.08	.08

GDS: Global Deficit Score

P-value[1]: comparisons between the HIV+ and HIV− samples

P-value[2]: comparisons between the reference and the validation samples

Table 2

Neuropsychological Test Battery

Verbal Fluency
Animal fluency * (Benton, Hamsher & Sivan, 1994; Heaton et al., 2004)
Letter fluency * (Benton, Hamsher & Sivan, 1994; Gladsjo et al., 1999)
Attention/Working Memory
PASAT-50 * (Gronwall, 1997; Diehr, Heaton & Miller, 1998)
WAIS-III L-N Sequencing (Wechsler Adult Intelligence Scale-Third Edition, Psychological Corporation, 1997; Heaton, Taylor & Manly, 2003)
Speed of Information Processing
WAIS-III Digit Symbol * (Wechsler Adult Intelligence Scale-Third Edition, Psychological Corporation, 1997, Heaton, Taylor & Manly, 2003)
WAIS-III Symbol Search (Wechsler Adult Intelligence Scale-Third Edition, Psychological Corporation, 1997, Heaton, Taylor & Manly, 2003)
Trail Making Test A (Army Individual Test Battery, 1944) *
Executive Functioning
WCST-64 (Kongs, Thompson, Iverson, & Heaton, 2000)
Trail Making Test B (Army Individual Test Battery, 1944) *
Learning/Memory
Verbal (Hopkins Verbal Learning Test-Revised) total learning & delayed recall (Brandt & Benedict, 2001)
Visual (Brief Visuospatial Memory Test-Revised) total learning & delayed recall (Benedict, 1997)
Motor
Grooved Pegboard dominant * & non-dominant hand * (Klove, 1963; Heaton et al., 2004)

* NP measures included in the abridged summary score for cross-validation.

Table 3

Rate of significant change in the HIV+ and HIV− reference samples, for the mean scaled score (90% confidence interval sRCS cut-offs), and continuous sRCS, results across study time points

	% decline		% stable		% improve			Continuous sRCS		
Times	**HIV− Reference Sample**	**HIV+ Reference Sample**	**HIV− Reference Sample**	**HIV+ Reference Sample**	**HIV− Reference Sample**	**HIV+ Reference Sample**	***p***	**HIV− Reference Sample**	**HIV+ Reference Sample**	***p***
Time 1 *pred* Time 2	4.65%	3.23%	90.70%	93.55%	4.65%	3.23%	.67	− 0.00 (1.00)	0.09 (0.90)	.39
Time 2 *pred* Time 3	5.13%	1.61%	89.74%	91.94%	5.13%	6.45%	.29	− 0.00 (1.00)	0.01 (0.95)	.94
Time 3 *pred* Time 4	4.30%	2.56%	91.40%	96.15%	4.30%	1.28%	.40	− 0.00 (1.00)	− 0.10 (0.95)	.50
Time 4 *pred* Time 5	4.23%	6.56%	91.55%	88.52%	4.23%	4.92%	.81	− 0.00 (1.00)	−0.02 (1.28)	.90

Summary regression-based change scores (sRCS)

Hierarchical models were developed in the 172 HIV− individuals only with previous performance in 1st order of entry and then demographic characteristics. The predictor's model at time 1 predicting time 2 included, the time 1 mean scaled score and age; the predictor's model at time 2 predicting time 3 included, the time 2 mean scaled score; the predictor's model at time 3 predicting time 4 included, the time 3 mean scaled score; the predictor's model at time 4 predicting time 5 included, the time 4 mean scaled score and education. The SRB equations were then built and applied to the HIV+ individuals.

Table 4

Baseline NP performance (scaled scores, mean (SD)) on 15 NP measures in the HIV− and HIV+ reference samples (separately and combined) and in the validation samples

	HIV− Reference Sample (N=172)	HIV+ Reference Sample (N=124)	Total Reference Sample (N=296)	HIV− Validation Sample (N=111)
Letter fluency (total correct)	10.61 (2.65)	9.72 (2.73)	10.24 (2.72)	11.23 (2.74)
Animal Fluency (total correct)	11.03 (2.34)	9.77 (2.51)	10.5 (2.49)	11.45 (2.56)
PASAT 50 (total correct)	9.76 (2.64)	8.27 (3.11)	9.13 (2.94)	10.24 (2.70)
WAIS-III L-N Sequencing (total correct)	10.39 (2.61)	9.45 (2.72)	9.98 (2.69)	-
WAIS-III Digit Symbol (total correct)	9.68 (2.69)	10.08 (2.89)	9.85 (2.78)	10.17 (2.66)
WAIS-III Symbol Search (total correct)	10.28 (2.64)	9.17 (3.05)	9.81 (2.87)	-
Trail Making Test A (time in seconds)	10.91 (2.32)	9.24 (2.25)	10.20 (2.43)	11.31 (2.69)
WCST-64 (perseverative errors)	9.00 (3.08)	7.07 (2.79)	8.19 (3.11)	-
Trail Making Test B (time in seconds)	10.99 (2.53)	9.64 (2.99)	10.42 (2.81)	11.63 (2.55)
HVLT-R Total Learning (total correct)	8.87 (2.71)	8.08 (2.92)	8.54 (2.83)	-
HVLT-R Delayed Recall (total correct)	8.96 (3.02)	7.60 (3.36)	8.39 (3.23)	-
BVMT Total Learning (total correct)	9.24 (2.93)	8.05 (2.92)	8.74 (2.98)	-
BVMT Delayed Recall (total correct)	9.07 (3.09)	7.84 (3.07)	8.56 (3.13)	-
Grooved Pegboard DH (time in seconds)	9.93 (2.67)	8.66 (2.88)	9.40 (2.82)	10.66 (2.28)
Grooved Pegboard NDH (time in seconds)	9.74 (2.61)	8.24 (2.94)	9.15 (2.78)	10.31 (2.40)

L-N: Letter-Number; WCST-64: Wisconsin Card Sorting Test; HVLT-R: Hopkins Verbal Learning Test-Revised; BVMT: Brief Visuospatial Memory Test-Revised; DH: Dominant hand; NDH, non-dominant hand

Table 5

Significant predictors (of change, $p<.10$) of follow-up performance on the 15 NP measures in the final hierarchical models (scaled scores)

	Previous score	Overall competence	Age	Education	Sex	Ethnicity	Test-retest interval	R^2 final model
Letter Fluency								
Time 1 *pred* time 2	+	+		+				.59
Time 2 *pred* time 3	+	+						.61
Time 3 *pred* time 4	+	+						.63
Time 4 *pred* time 5	+	+						.58
Time 1 *pred* time 3	+							.60
Time 1 *pred* time 4	+				−			.63
Time 1 *pred* time 5	+							.49
Animal Fluency								
Time 1 *pred* time 2	+	+						.48
Time 2 *pred* time 3	+	+						.46
Time 3 *pred* time 4	+	+						.53
Time 4 *pred* time 5	+	+		+				.50
Time 1 *pred* time 3	+	+						.33
Time 1 *pred* time 4	+	+	−					.43
Time 1 *pred* time 5	+	+				+		.46
PASAT 50								
Time 1 *pred* time 2	+	+	−	+				.68
Time 2 *pred* time 3	+	+			−			.71
Time 3 *pred* time 4	+	+				+		.73
Time 4 *pred* time 5	+	+			−			.72
Time 1 *pred* time 3	+	+	−		−			.65
Time 1 *pred* time 4	+	+	−					.61
Time 1 *pred* time 5	+	+	−		−			.60
WAIS III L-N sequencing								
Time 1 *pred* time 2	+	+				+		.45

	Previous score	Overall competence	Age	Education	Sex	Ethnicity	Test-retest interval	R^2 final model
Time 2 *pred* time 3	+	+			−			.52
Time 3 *pred* time 4	+	+				+		.45
Time 4 *pred* time 5	+	+						.42
Time 1 *pred* time 3	+	+			−			.45
Time 1 *pred* time 4	+	+			−			.50
Time 1 *pred* time 5	+	+			−			.44
WAIS III Digit symbol								
Time 1 *pred* time 2	+	+	−		+			.70
Time 2 *pred* time 3	+	+	−					.74
Time 3 *pred* time 4	+	+				+		.80
Time 4 *pred* time 5	+	+		+				.84
Time 1 *pred* time 3	+	+			+			.68
Time 1 *pred* time 4	+	+						.67
Time 1 *pred* time 5	+	+	−					.70
WAIS III Symbol Search								
Time 1 *pred* time 2	+	+						.59
Time 2 *pred* time 3	+	+						.68
Time 3 *pred* time 4	+	+		+				.69
Time 4 *pred* time 5	+	+						.65
Time 1 *pred* time 3	+	+	−					.64
Time 1 *pred* time 4	+	+						.59
Time 1 *pred* time 5	+	+						.57
Trail Making Test A								
Time 1 *pred* time 2	+	+	−					.50
Time 2 *pred* time 3	+	+	−			+		.50
Time 3 *pred* time 4	+	+		+				.57
Time 4 *pred* time 5	+	+	−					.52
Time 1 *pred* time 3	+	+	−			+		.46
Time 1 *pred* time 4	+	+	−	+				.51

	Previous score	Overall competence	Age	Education	Sex	Ethnicity	Test-retest interval	R^2 final model
Time 1 *pred* time 5	+	+	−					.46
Trail Making Test B								
Time 1 *pred* time 2	+	+	−					.62
Time 2 *pred* time 3	+	+				+		.67
Time 3 *pred* time 4	+	+	−					.69
Time 4 *pred* time 5	+	+	−					.71
Time 1 *pred* time 3	+	+	−			+		.66
Time 1 *pred* time 4	+	+	−					.64
Time 1 *pred* time 5	+	+	−					.61
WCST Perseverative errors								
Time 1 *pred* time 2	+	+	−			+		.31
Time 2 *pred* time 3	+	+						.45
Time 3 *pred* time 4	+	+	−					.51
Time 4 *pred* time 5	+	+		+				.46
Time 1 *pred* time 3	+	+	−			+		.42
Time 1 *pred* time 4	+	+	−					.45
Time 1 *pred* time 5	+	+	−					.30
HVLT-R Total Learning								
Time 1 *pred* time 2	+	+	−					.49
Time 2 *pred* time 3	+	+						.40
Time 3 *pred* time 4	+	+						.46
Time 4 *pred* time 5	+	+		+				.58
Time 1 *pred* time 3	+	+					−	.42
Time 1 *pred* time 4	+	+					−	.37
Time 1 *pred* time 5	+	+		+				.49
HVLT-R Delayed Recall								
Time 1 *pred* time 2	+	+	−					.43
Time 2 *pred* time 3	+	+						.38

	Previous score	Overall competence	Age	Education	Sex	Ethnicity	Test-retest interval	R^2 final model
Time 3 *pred* time 4	+	+		+				.43
Time 4 *pred* time 5	+	+						.49
Time 1 *pred* time 3	+	+	−					.39
Time 1 *pred* time 4	+	+					−	.32
Time 1 *pred* time 5	+	+						.31
BVMT Total Learning								
Time 1 *pred* time 2	+	+	−	+				.49
Time 2 *pred* time 3	+	+						.42
Time 3 *pred* time 4	+	+						.49
Time 4 *pred* time 5	+	+		+				.45
Time 1 *pred* time 3	+	+	−					.45
Time 1 *pred* time 4	+	+	−				−	.49
Time 1 *pred* time 5	+	+		+	−			.46
BVMT Delayed Recall								
Time 1 *pred* time 2	+	+	−					.39
Time 2 *pred* time 3	+	+						.29
Time 3 *pred* time 4	+	+						.33
Time 4 *pred* time 5	+	+	−				−	.33
Time 1 *pred* time 3	+	+					−	.36
Time 1 *pred* time 4	+	+					−	.29
Time 1 *pred* time 5	+	+	−					.28
Grooved Peg DH								
Time 1 *pred* time 2	+	+	−					.57
Time 2 *pred* time 3	+	+	−					.59
Time 3 *pred* time 4	+	+						.56
Time 4 *pred* time 5	+	+						.60
Time 1 *pred* time 3	+	+	−					.52
Time 1 *pred* time 4	+	+	−		+			.53
Time 1 *pred* time 5	+	+	−					.52

	Previous score	Overall competence	Age	Education	Sex	Ethnicity	Test-retest interval	R^2 final model
Grooved Peg NDH								
Time 1 *pred* time 2	+	+	−		+			.56
Time 2 *pred* time 3	+	+	−					.61
Time 3 *pred* time 4	+	+						.63
Time 4 *pred* time 5	+	+						.63
Time 1 *pred* time 3	+	+	−					.57
Time 1 *pred* time 4	+	+	−					.52
Time 1 *pred* time 5	+	+	−					.48

pred: predicting

Based on effect in overall model: (+) Positive correlation between predictor and outcome; (−) Negative correlation between predictor and outcome.

Hierarchical models: order of analysis: 1. Previous test score; 2. overall competence at previous time; 3. age; 4. education; 5. Sex (male 0; female 1); 6 ethnicity (white 1, other 0); 7. Test-retest interval (months)

L-N: Letter-Number; WCST-64: Wisconsin Card Sorting Test; HVLT-R: Hopkins Verbal Learning Test-Revised; BVMT: Brief Visuospatial Memory Test-Revised; DH: Dominant hand; NDH, non-dominant hand

Table 6

(Median) practice effect from baseline to follow-up on 15 NP measures (scaled scores)

	T2	T3	T4	T5 +
Letter fluency	0.0	0.5	1.0	1.0
Animal Fluency	0.0	0.0	0.0	0.0
PASAT 50	0.5	1.0	1.0	1.0
WAIS-III L-N Sequencing	0.0	0.0	0.0	0.0
WAIS-III Digit Symbol	0.0	0 .5	1.0	1.0
WAIS-III Symbol Search	0 .5	1.0	1.0	1.0
Trail Making Test A	0.5	1.0	1.0	1.0
WCST-64 perseverative errors	1.0	2.0	2.0	2.0
Trail Making Test B	1.0	1.0	1.0	1.0
HVLT-R Total Learning	0.0	1.0	0 .5	0 .5
HVLT-R Delayed Recall	0 .5	0 .5	0 .5	0 .5
BVMT Total Learning	1.0	1.0	0.0	1.0
BVMT Delayed Recall	0 .5	0.0	0.0	0 .5
Grooved Pegboard DH	0.5	0.0	1.0	1.0
Grooved Pegboard NDH	0.0	0 .5	0.5	1.0
Sum	6.0	10.0	10.5	12.5

L-N: Letter-Number; WCST-64: Wisconsin Card Sorting Test; HVLT-R: Hopkins Verbal Learning Test-Revised; BVMT: Brief Visuospatial Memory Test-Revised; DH: Dominant hand; NDH, non-dominant hand

Table 7

Uncorrected and practice-effect corrected Mean scaled score across reference samples and across study time points

	T1	T2	T3	T4	T5	T6
2 testings, N=296	9.40 (1.78)	9.85 (1.87)	-	-	-	-
	-	9.45 (1.87)				
3 testings, N=241	9.30 (1.81)	9.76 (1.84)	9.90 (1.91)	-	-	-
	-	9.36 (1.84)	9.24 (1.91)			
4 testings, N=171	9.35 (1.77)	9.81 (1.79)	9.95 (1.88)	9.97 (1.96)	-	-
	-	9.41 (1.79)	9.28 (1.88)	9.27 (1.96)		
5 testings, N=132	9.54 (1.79)	9.96 (1.77)	10.12 (1.89)	10.20 (1.89)	10.27 (2.00)	-
	-	9.56 (1.77)	9.45 (1.89)	9.50 (1.89)	9.43 (2.01)	
6 testings, N=64	9.32 (1.81)	9.74 (1.78)	9.86 (1.82)	9.94 (1.76)	9.98 (1.90)	10.26 (1.98)
	-	9.34 (1.78)	9.20 (1.81)	9.24 (1.76)	9.15 (1.91)	9.36 (1.98)

Table 8

Reference sample reliability statistics (r) on the 15 NP measures (scaled scores) across study time points.

Tests	T1 * T2 N=296	T2 * T3 N=241	T3 * T4 N=171	T4 * T5 N=132	T5*T6 N=64
Mean Scaled Score	.91	.91	.92	.91	.88
Letter Fluency	.76	.77	.78	.75	.75
Animal Fluency	.65	.66	.70	.66	.68
PASAT 50	.77	.82	.85	.84	.82
WAIS-III L-N Sequencing	.65	.68	.62	.61	.59
WAIS-III Digit Symbol	.82	.84	.89	.91	.92
WAIS-III Symbol Search	.73	.78	.82	.79	.82
TMT A	.67	.61	.70	.64	.73
TMT B	.73	.79	.82	.80	.74
WCST perseverative errors	.43	.56	.66	.64	.67
HVLT-R Total Learning	.67	.60	.65	69	.53
HVLT-R Delayed Recall	.63	.57	.62	.67	.60
BVMT Total Learning	.63	.63	.65	.64	.67
BVMT Delayed Recall	.55	.48	.52	.49	.52
Grooved Pegboard DH	.74	.74	.72	.77	.83
Grooved Pegboard NDH	.71	.76	.79	.78	.57

TMT: Trail Making Test; L-N: Letter-Number; WCST: Wisconsin Card Sorting Test; HVLT-R: Hopkins Verbal Learning Test-Revised; BVMT: Brief Visuospatial Memory Test-Revised; DH: Dominant hand; NDH, non-dominant hand.

NB: Alternate forms were used at each follow-up sessions for the memory tests: Hopkins Verbal Learning Test-Revised; BVMT: Brief Visuospatial Memory Test-Revised

Table 9

Performance of abridged sRCS in the norms development (reference sample versus validation sample)

	% decline		% stable		% improve		Continuous sRCS	
Times	**Reference Sample**	**Validation Sample**	**Reference Sample**	**Validation Sample**	**Reference Sample**	**Validation Sample**	**Reference Sample**	**Validation Sample**
Time 1 *pred* Time 2	5%	6%	90%	93%	5%	1%	−0.0004 (0.46)	−0.16 (0.42)
Time 2 *pred* Time 3	5%	7%	90%	90%	5%	3%	0.0007 (0.47)	−0.04 (0.45)
Time 1 *pred* Time 3	5%	6%	90%	91%	5%	3%	0.0000 (0.51)	−0.12 (0.45)

sRCS: summary regression based change score (in these analyses the sRCS is abridged including 8 NP measures: Grooved Peg Board dominant hand and non dominant hand; Trail Making Test A & B; Digit Symbol, PASAT 50; Letter and Semantic Fluency)

pred: predicting

N=111 (cross validation samples with 2 visits): comparisons with the reference sample showed that the cross-validation sample tended to decline more on the continuous sRCS at time 2. However, there were no significant differences when using the discrete classifications (decline/improve/stable; p=.15).

N= 67 (cross-validation samples with 3 visits): comparisons with the reference sample showed that the cross-validation sample tended to decline more on the continuous sRCS continuous sRCS at time 3, but this was not significant (p=.52). Also there were no significant differences when using the discrete classifications (decline/improve/stable; p-values ranging between .22 and .75).