Published in final edited form as:

J Clin Exp Neuropsychol. 2007 August ; 29(6): 669–678. doi:10.1080/13803390600920471.

The generalizability of neurocognitive test/retest data derived from a nonclinical sample for detecting change among two HIV+ cohorts

Andrew J. Levine¹, Charles H. Hinkin², Eric N. Miller^{2,3}, James T. Becker⁴, Ola A. Selnes⁵, and Bruce A. Cohen⁶ for the Multicenter AIDS Cohort Study (MACS)

¹ Department of Neurology, National Neurological AIDS Bank, UCLA–David Geffen School of Medicine, Los Angeles, CA, USA

² Department of Psychiatry and Biobehavioral Sciences, UCLA–David Geffen School of Medicine, Los Angeles, CA, USA

³ Department of Psychology, UCLA, Los Angeles, CA, USA

⁴ Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

⁵ Department of Neurology, Johns Hopkins Medical Institutions, Baltimore, MD, USA

⁶ Department of Neurology, Northwestern University, Chicago, IL, USA

Abstract

Objective methods for determining clinically relevant neurocognitive change are useful for clinicians and researchers, but the utility of such methods requires validation studies in order to assess their accuracy among target populations. We examined the generalizability of regression equations and reliable change indexes (RCI) derived from a healthy sample to two HIV-infected samples, one similar in demographic makeup to the normative group and the other dissimilar. Measures administered at baseline and follow-up included the Trail Making Test, Controlled Oral Word Association Test (COWAT), Grooved Pegboard, and Digit Span. Frequencies of decline, improvement, or stability were determined for each measure. Among the demographically similar clinical cohort, elevated rates of decline among more immunologically impaired participants were indicated by simple regression method on measures of psychomotor speed and attention, while RCI addressing practice effects (RCI-PE) indicated improvement on most measures regardless of immunostatus. Conversely, among the demographically dissimilar cohort, simple regression indicated high rates of decline across all measures, while RCI-PE indicated elevated rates of decline on psychomotor and attention measures. Thus, the accuracy of the two methods examined for

Address correspondence to Andrew Levine, 11645 Wilshire Blvd., Suite 770, Los Angeles, CA 90025, USA (ajlevine@med-net.ucla.edu).

Publisher's Disclaimer: Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) with centers (Principal Investigators) at The Johns Hopkins University Bloomberg School of Public Health (Joseph B. Margolick, Lisa Jacobson), Howard Brown Health Center and Northwestern University Medical School (John Phair), University of California, Los Angeles (Roger Detels), and University of Pittsburgh (Charles Rinaldo) (Website located at <http://www.statepi.jhsph.edu/mac/mac.html>).

determining clinically significant change among HIV+ cohorts differs depending upon their similarity with the normative sample.

Detecting change in neurocognitive functioning is one of the primary roles of the neuropsychologist. This is often done without the benefit of baseline data; however, even when such information is available it is not always possible to determine the clinical significance of change in test scores. For example, changes in scores can be due to poor reliability of the measure, regression towards the mean, or intraindividual variations in behavior and motivation. However, even when these factors are considered, improved performance still may not represent clinically meaningful change. Perhaps the greatest obstacle in interpreting retest data is the effect that prior exposure to the test has on performance, commonly referred to as “practice effect.” This is most evident in settings in which changes in ability are expected and need to be quantified to some extent, such as neurorehabilitation, neurosurgery, or the assessment of neurodegenerative diseases. However, despite the relevance of detecting change to the practice of neuropsychology, relatively few studies exist that provide useful data for interpreting retest performance or that compare the various statistical methods for determining change.

There are a number of statistical approaches available for controlling for practice effects and therefore assisting in determining relevance change (for review, see Collie, Darby, Falletti, Silbert, & Maruff, 2002). These include the reliable change index (RCI) and its variants that address practice effects (RCI-PE; Chelune, Naugle, Luders, Sedlak, & Awad, 1993), simple regression, and multiple regression. The last two have the additional advantage of being able to correct for regression towards the mean. Additionally, multiple regression is capable of considering additional factors that might affect one’s capacity to benefit from prior exposure to a measure, such as age, education, and test–retest interval. However, while multiple regression accounts for the variances that these factors contribute, they are generally small and are seen only on a few measures (Levine, Miller, Becker, Selnes, & Cohen, 2004; Salinsky, Storzbach, Dodrill, & Binder, 2001; Temkin, Heaton, Grant, & Dikmen, 1999). Therefore, the relative simplicity of simple regression and RCI methods outweighs the small benefit of multiple regression in most instances.

Whatever technique one decides to use, the data must be derived from a control sample. Herein lies another area of debate: whether to use nonclinical or similar clinical samples in developing the RCI and regression formulas to be used with the sample of interest.

Perhaps the greatest contribution in neuropsychological test–retest literature comes from epilepsy researchers interested in the outcomes of temporal resection surgery. The trend among these researchers has been to use clinical samples from which to derive their data: that is, individuals with epilepsy who do not undergo surgical intervention. Earlier papers provided test–retest norms derived for a wide range of commonly used neuropsychological measures (Hermann et al., 1996; Sawrie, Chelune, Naugle, & Luders, 1996). More recently, others have reported such data for batteries such as the Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Sherman et al., 2003), Wechsler Adult Intelligence Scale–Third Edition (WAIS-III; Basso, Carona, Lowery, & Axelrod, 2002), and Wechsler Memory Scale–Third Edition (WMS-III; Martin et al., 2002). Across these studies, the success of the data in detecting change as a result of surgery has been good, at least as an adjunctive method.

Others have used healthy cohorts to derive data that might be applied to clinical populations for the purpose of filtering out the effects of practice and detecting clinically relevant change (Dikmen, Heaton, Grant, & Temkin, 1999; Ivnik et al., 1999; Levine et al., 2004; Salinsky et al., 2001; Temkin et al., 1999). However, little information is available to answer whether or not data derived from healthy, nonclinical cohorts are applicable to clinical patients. Findings

by Heaton et al. (2001) demonstrated that norms derived from a nonclinical group are not particularly useful for predicting retest performance among clinical populations. Those authors compared the efficacy of RCI-PE and simple and multiple regression in determining change in two groups of neurologically stable individuals (a nonclinical cross-validation group and a group diagnosed with schizophrenia) and two groups of neurologically unstable individuals (a sample recovering from traumatic brain injury and a group that had incurred brain insult between baseline and follow-up testing, called the “recovering TBI” and “new insult” group, respectively). Among the two stable groups, all three methods performed equally well with regard to specificity in classifying subjects as unchanged at retest. However, large standard errors obtained on some measures among the schizophrenic participants, with the result of an excessive number being classified as “changed” in both the positive and negative directions, led the authors to conclude that confidence intervals developed from the nonclinical sample may not be appropriate for such psychiatrically impaired individuals. Among the unstable groups the three methods were generally equivalent with regards to sensitivity, or the ability to detect change when it did occur. The authors concluded that it would be wise to use norms derived from individuals who are similar in demographics and baseline performance to the population to whom the data will be applied. More recently, this same group followed their suggestion when they examined the accuracy of modified RCI equations derived from a healthy sample in determining change on a neuropsychological test battery among a demographically similar HIV+ cohort (Woods et al., 2006). The equations demonstrated good specificity for the HIV+ participants after a one-year test–retest interval.

Recently, Levine et al. (2004) reported test–retest data and regression equations for eight neuropsychological measures that were derived from 1,047 healthy (HIV-negative) individuals. The sample consisted primarily of well-educated, Caucasian men. However, as the results of Heaton et al. (2001) suggest, the generalizability of data derived from that sample may not extend to clinical groups. In the current study, we examined the utility of the test–retest data derived from that sample in detecting change among two HIV-positive groups: one a demographically similar sample and the other demographically different. Individuals with more advanced disease (CD4+ T-cells < 200) were expected to show higher rates of clinically relevant decline on measures most sensitive to HIV-related neuropsychological deficits. We also expected that the data would be more accurate in predicting changed performance among individuals in the demographically similar group who were more immunocompromised, while the demographically dissimilar cohort would show unusually high rates of change.

METHOD

Research participants

Three groups comprised the total study sample. Two were drawn from the Multicenter AIDS Cohort Study (MACS), database. Briefly, the MACS is a multicenter epidemiological study of the natural history of HIV infection, conducted in four U.S. cities (Baltimore, Chicago, Pittsburgh, Los Angeles). Recruitment procedures have been described elsewhere (Kaslow et al., 1987, Miller et al., 1990). Participants were generally evaluated at semiannual intervals, although test–retest intervals varied in length (see Tables 1 and 3). The original healthy sample from which retest data was derived had a test–retest interval of between 4 and 24 months. Evaluations included physical examinations, HIV testing, structured clinical interviews, and neuropsychological testing. MACS participants were excluded from the analyses if they had a self-reported history of significant head trauma (loss of consciousness greater than 1 hour), self-reported history of learning disability, and/or substance abuse in the 6-month period prior to initial testing (cocaine, heroin, PCP, or methamphetamines).

The first group, from which the regression equations and RCI-PE data were derived, was a non-clinical sample drawn from a pool of 1,047 healthy, HIV-negative males from the MACS

database. This sample was described in detail in Levine et al. (2004). Mean age and education for this group varied across neuropsychological measure and are listed in Table 1. The ethnic makeup was 87.7% Caucasian, 7.1% Black, 3.7% Hispanic, and the remaining 1.5% Pacific Islander, Native American, or Asian.

The second group was a demographically similar clinical (HIV-seropositive) cohort drawn from a pool of 955 MACS participants. Overall demographic characteristics of this group are listed in Table 2. There were no females in the MACS study and no one with current substance dependence as defined by the Diagnostic and Statistical Manual Of Mental Disorders (DSM-IV) criteria (American Psychiatric Association, 1994). A total of 57% met diagnostic criteria for acquired immunodeficiency syndrome (AIDS). Due to variation in the battery administered among study sites, the characteristics of the participants varied across neuropsychological measure (see Table 3). The overall ethnic makeup of this group was 81.6% Caucasian, 10.4% Black, 6.6% Hispanic, and the remaining 1.4% Pacific Islander, Native American, or Asian.

The third group consisted of 173 HIV-positive individuals who were participating in a 6-month, National Institute of Drug Abuse (NIDA)-funded study of antiretroviral medication adherence among drug users. Participants were recruited from the Los Angeles area through advertisements posted at university-affiliated infectious disease clinics as well as through community-based HIV/AIDS organizations. A total of 61 (65%) of the participants in his group met diagnostic criteria for AIDS. A total of 28 (16%) were female. A total of 55 (32%) of the participants were classified as currently drug dependent at the second visit according to DSM-IV criteria (American Psychiatric Association, 1994), as determined through the Psychiatric Research Interview for Substance and Mental Disorders (PRISM; Hasin et al., 1996). As with the MACS cohort, NIDA participants were excluded from the analyses if they had a history of learning disability or head trauma with loss of consciousness greater than 1 hour. Demographic characteristics of this group are listed in Table 4.

Both clinical cohorts were further divided into groups based on Centers for Disease Control (CDC) staging based on CD4+ T-cells count. Thus, individuals in each cohort were classified as follows: Group 1 (CD4+ T-cells < 200), Group 2 (CD4+ T-cells = 200–499), Group 3 (CD4+ T-cells = 500+).

Measures

The following measures were used: Trail Making Test–Parts A and B (Army Individual Test Battery, 1944); Stroop Color/Word Interference Test (Kaplan adaptation of the Comalli administration procedure; Stroop, 1935; Comalli, Wapner, & Werner, 1962); Grooved Pegboard (Klove, 1963; Lafayette Instrument Company); Digit Span (from the Wechsler Adult Intelligence Scale-Revised; Wechsler, 1981); Symbol Digit Modalities Test (Smith, 1991); and Controlled Oral Word Association Test (COWAT; Benton & Hamsher, 1978).

Data analysis

Raw scores were used for all analyses. Derivation of the original regression equations and data for calculating RCIs are detailed in a previously published paper (Levine et al., 2004). Briefly, regression equations were created by entering scores from the healthy normative sample into a simple regression analysis, with baseline scores as the predictor and retest scores as the criterion. Thus, β -coefficients and residual standard deviations were derived from these former analyses for each measure. These can be found in Levine et al., 2004. In the current analyses, baseline scores for the clinical groups were used as the predictor variable, with the retest score as the dependent variable. The residual standard deviations from the healthy group were used to create confidence intervals (CIs) around the predicted scores. The equation is as follows:

$$\frac{(X_2 - \text{predicted } X_2)}{\text{residual } SD} \quad (1)$$

where: $X_2 - \text{predicted } X_2$ is the difference in between actual retest score and predicted retests score, and residual SD is the residual standard deviation of the normative sample.

Using the RCI-PE method, difference scores were derived from the normative sample in order to determine significance of change in performance. The RCI-PE (Chelune et al., 1993) is calculated by subtracting the mean change score (retest – baseline) of a healthy normative sample from the individual's or group of interest's mean change score. This is then divided by the standard deviation of the mean difference scores, or standard error of measurement of the difference (SD_{diff}), of the normative sample (data provided in Levine et al., 2004). The equation for calculating the RCI-PE is as follows:

$$\frac{(X_2 - X_1) - (M_2 - M_1)}{SD_{\text{diff}}} \quad (2)$$

where: $X_2 - X_1$ is the difference in between an individual's or research sample's baseline and retest performance; $M_2 - M_1$ is the difference between the normative sample's mean baseline and retest performances; SD_{diff} is the standard deviation of difference between baseline and retest scores of the normative sample.

In the current study, a 90% CI was constructed around the RCI-PE and regression estimates. Therefore, among the nonclinical sample, 90% of the participants should have scores that fall between 1.645 and -1.645, with 5% being in the positive and 5% in the negative directions.

Using the confidence intervals described above, frequencies of declined, improved, and stable performances were determined for each measure among the three CDC-defined groups using both the regression equations and RCI-PE.

RESULTS

MACS HIV+ cohort

Baseline and retest scores are displayed in Table 5. As shown, there was ubiquitous improvement on all tests, reaching statistical significance in all instances. Frequencies of change on neuropsychological measures, as determined via the regression and RCI-PE, are presented in Table 6. Results of the RCI-PE equations indicate greater rates of improvement on Trail Making Test (Form A) and Grooved Pegboard (both hands). Further, this improvement appears to be similar across CDC groups, although slightly less for the most immunologically compromised (i.e., CD4+ T-cells < 200). Only on the COWAT did the RCI-PE equations indicate a decline rate of greater than 10%, and this was similar across CDC groups. Conversely, rates of decline and improvement consistently fell below 10% using the regression method, with the exception of Trail Making Test (Form B). On that measure, an expected higher rate of decline was seen among the more immunologically compromised. Note that there was a general incongruence in direction of change between the two methods, with regression showing somewhat higher rates of decline (5.7–11.3%) and RCI-PE indicating higher rates of improvement (3.9–16.1%).

NIDA HIV+ cohort

Baseline and retest scores are shown in Table 5. As with the MACS cohort, all tests showed statistically significant change from baseline. However, in the NIDA sample, scores on both trials of the Grooved Pegboard and on Digit Span were poorer at retest. Rates of change for this sample are shown in Table 7. Elevated (>10%) rates of decline based on RCI-PE appeared on Trail Making Test (Form B) and on Grooved Pegboard (both hands). However, rates of decline did not appear to be associated with CDC group for any of these measures.

Improvement rates were also elevated for Trail Making Test (Form B); however, once again there appeared to be no relation with CDC group. Rates of change based on regression were markedly higher than those based on RCI-PE, and exclusively in the direction of decline. This included a decline rate of 22.1% on Trail Making Test Form A, 35% on Form B, 40.2% on Grooved Pegboard dominant hand, 42.3% on nondominant hand, 19.1% on the COWAT, and 12.3% on Digit Span. As with the RCI-PE method, little association between decline rates and CDC group was apparent.

DISCUSSION

In this study, we assessed the generalizability of data derived from a healthy sample in detecting clinically relevant change among two groups: HIV-infected individuals who were demographically similar to the healthy sample and a group of HIV-infected individuals who were predominantly drug users and demographically dissimilar to the healthy sample. Development and validation of test–retest data is extremely valuable to neuropsychologists and other healthcare professionals as a means of assessing the meaningfulness of change over time. This study is among the few that addresses the generalizability of test–retest data.

When applied to a clinical cohort with similar demographic characteristics, the pattern of changes indicated by the regression equations was consistent with neurocognitive changes that one would expect in HIV-infected individuals—that is, increasing rates of decline on tests of psychomotor speed, working memory, and complex attention were associated with greater virologic compromise (Heaton et al., 1995). Conversely, the RCI-PE method indicated somewhat greater rates of improvement on these tasks, which was similar regardless of immunological status. Which method is more accurate is impossible to determine from our dataset, as no objective measure of change was established to use as a criterion. However, because the bulk of the data were collected prior to widespread use of highly active antiretroviral therapy (HAART), it would be expected that the sample was more prone to decline than improvement between baseline and retest. Thus, it appears that regression equations derived from a healthy normative sample do generalize well to a demographically similar clinical cohort, while RCI-PE is less accurate.

Conversely, among a clinical cohort that was quite different from the normative sample with regards to ethnicity, education, and other factors (e.g., substance abuse rates), it was the RCI-PE data that appear to have provided a more accurate assessment of change. Among this cohort, the regression equations predicted an unusually high rate of declined performance across all measures. This finding is somewhat understandable in the cases of the Grooved Pegboard and Digit Span, on which this cohort actually did perform worse as a group at retest. However, it is unclear why such high rates of decline were seen on the other measures. One possible reason for the apparent poor generalizability of regression equations to the NIDA sample was that there was in fact a high rate of decline among that cohort. This was a cohort that consisted largely of cocaine and methamphetamine abusers, drugs that have been shown to have an additive impact upon cognitive functioning in those with HIV (Levine et al., 2006; Rippeth et al., 2004). Therefore, the higher rates of decline indicated by the regression equations may be an accurate reflection of actual advancing cognitive deficit over the course of the 6-month study. However, examination of the group mean test scores and standard deviations (see Table

5) suggests that significant decline may have occurred in a subset of individuals, but that the group as a whole generally improved or maintained initial levels of performance across tests. As pointed out by Dikmen et al. (1999), individuals with low initial scores demonstrate the greatest change at retest due to a number of factors, including practice effects and regression to their true ability level. However, the opposite direction of change was observed in this cohort of individuals who performed worse as a group than did the other clinical cohort. In contrast to regression, when the RCI-PE data were applied to the NIDA cohort, rates of change were more modest. Only on the Grooved Pegboard (both hands) and Trail Making Test–Form B were elevated rates of decline seen, albeit considerably more modest than those of the regression method. Finally, there appears to be little relationship between rates of decline on these measures and immunostatus.

The reason for the vastly different findings based on regression versus RCI-PE may lie in the inherent characteristics of the two methods. In our NIDA sample, regression-based change formulas indicated high rates of clinically significant decline on all measures, including those in which there was an overall group improvement. Authors of previous studies comparing various methods for determining clinically significant change reported that simple models, such as the regression method used here, may be appropriate only when used with individuals who have typical baseline performance and who are homogeneous demographically (Temkin et al., 1999). Clearly, our NIDA sample did not meet these criteria. Those authors also suggested that wider confidence intervals be used for those with poor baseline performance in order to increase specificity and narrower confidence intervals for those with those with normal baseline performance in order to improve sensitivity. In many of our participants, lack of improvement at retest may have resulted in classification as “declined,” as their scores fell below the cutoff predicted by the residual term. Thus, wider confidence intervals may have been more appropriate for individuals who have atypical baseline scores. This inherent limitation of standard regression equations has been discussed in detail by Crawford and Howell (1998), who showed that regression equations derived from small-to-moderate sized samples ($N < 100$) tend to have confidence intervals that are too narrow to accurately classify individuals from a population of interest. For large samples, such as that used in the current study, this is not an issue. However, those authors showed that extreme scores in the population of interest can also result in erroneous change classification (i.e., declined or improved), as appears to be the case with our NIDA cohort. Thus, inflated Type I error rates can be expected when regression equations obtained from a healthy sample are applied to a sample with widely varying scores, as demonstrated in the NIDA sample of our study. This was not seen when regression was applied to our MACS clinical sample, as scores tended to be more centrally distributed. Crawford and Howell (1998), and more recently Crawford and Garthwaite (2006), have proposed a more accurate, albeit complicated, method for determining regression formulas and confidence intervals from a healthy population that could lead to greater accuracy in determining change. A simple rule of thumb, based on previous studies and the current findings, is that simple regression may be the more appropriate method when the sample of interest is demographically similar to the normative sample and is relatively large ($N > 100$) or has somewhat homogeneous baseline scores. When the sample of interest is demographically dissimilar to the normative sample, RCI-PE appears to be the better method for determining clinically relevant change.

We acknowledge that the absolute amount of improvement or decline, as reflected by change in scores, was small across measures. For example, among the MACS group, performance on the COWAT improved by only a single word. However, this was statistically significant, presumably due to little variation in direction and degree of change across individuals in this group. This small degree of change was comparable to that of the normative sample (see Levine et al., 2004) across all measures, suggesting that it does indeed reflect more than random variation or regression to the mean. Such findings lend support for the validity of these

neuropsychological tests in detecting true change, at least on a group level. The true utility of the RCI and regression methods in determining significance of change will continue to be elucidated in studies that employ a criterion measure, such as collateral ratings, neuroimaging, and clinician diagnoses that do not consider neuropsychological test performance.

Finally, it is interesting that RCI-PE, which does not correct for variability in baseline performance to the extent that regression does, resulted in seemingly more expected change rates among the NIDA sample. Even within the MACS sample, rates of decline based on regression were higher than those based on RCI-PE for Grooved Pegboard and Trail Making Tests. Again, whether this indicates lack of improvement due to true pathology among a subset of individuals or poor specificity of the regression-based formula is unclear. Further research, in which a concurrent measure of neurocognitive change is available (e.g., clinical rating or neurologic diagnosis), is needed to probe this issue.

Acknowledgments

The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute and the National Heart, Lung and Blood Institute. UO1-AI-35042, 5-MO1-RR-00722 (GCRC), UO1-AI-35043, UO1-AI-37984, UO1-AI-35039, UO1-AI-35040, UO1-AI-37613, UO1-AI-35041. This study was also supported in part by a grant from NIDA (RO1 DA13799) and data provided by Charles Hinkin, PhD.

References

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 1994.
- Army Individual Test Battery. Manual of directions and scoring. Washington, DC: War Department, Adjutant General's Office; 1944.
- Basso M, Carona F, Lowery N, Axelrod B. Practice effects on the WAIS-III across 3 and 6-month intervals. *The Clinical Neuropsychologist* 2002;16:57–63. [PubMed: 11992227]
- Beck, A. Beck Depression Inventory. San Antonio, TX: The Psychological Corporation; 1987.
- Beck, A.; Steer, R.; Brown, G. Beck Depression Inventory–Second Edition manual. San Antonio, TX: The Psychological Corporation; 1996.
- Benton, A.; Hamsher, K. Multilingual Aphasia Examination. Iowa City, IA: University of Iowa; 1978.
- Chelune G, Naugle R, Luders H, Sedlak J, Awad I. Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology* 1993;7:41–52.
- Collie A, Darby D, Falletti M, Silbert B, Maruff P. Determining the extent of cognitive change after coronary surgery: A review of statistical procedures. *Annals of Thoracic Surgery* 2002;73:2005–2011. [PubMed: 12078822]
- Comalli PE, Wapner S, Werner H. Interference effects of Stroop Color-Word Test in childhood, adulthood, and aging. *Journal of Genetic Psychology* 1962;100:47–53. [PubMed: 13880724]
- Crawford JR, Garthwaite PH. Comparing patients predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence intervals. *Neuropsychology* 2006;20:259–271. [PubMed: 16719619]
- Crawford JR, Howell DC. Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology* 1998;20:755–762. [PubMed: 10079050]
- Dikmen S, Heaton R, Grant I, Temkin N. Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society* 1999;5:346–356. [PubMed: 10349297]
- Grober E, Sliwinski M. Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology* 1991;13:933–949. [PubMed: 1779032]

- Heaton R, Grant I, Butters N, White DA, Kirson D, et al. The HNRC 500—neuropsychology of HIV infection at difference stages: The HIV Neurobehavioral Research Center. *Journal of the International Neuropsychological Society* 1995;1:231–251. [PubMed: 9375218]
- Heaton R, Temkin N, Dikmen S, Avitable N, Taylor M, et al. Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology* 2001;16:75–91. [PubMed: 14590193]
- Hermann B, Seidenberg M, Schoenfeld J, Peterson J, Leveroni C, Wyler A. Empirical techniques for determining reliability, magnitude, and pattern of neuropsychological change after epilepsy surgery. *Epilepsia* 1996;37:942–950. [PubMed: 8822692]
- Ivnik R, Smith G, Lucas J, Petersen R, Boeve B, Kokmen E, et al. Testing normal older people three or four times at 1-to-2 year intervals: Defining normal variance. *Neuropsychology* 1999;13:121–127. [PubMed: 10067783]
- Kaslow R, Ostrow D, Detels R, Phair J, Polk B, Rinaldo C. The Multicenter AIDS Cohort Study: Rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology* 1987;26:310–318. [PubMed: 3300281]
- Klove, H. Clinical neuropsychology. In: Forster, FM., editor. *Medical Clinics of North America*. New York: Saunders; 1963.
- Levine A, Hardy D, Miller E, Castellon S, Longshore D, Hinkin C. The effect of recent stimulant use on sustained attention in HIV-infected adults. *Journal of Clinical and Experimental Neuropsychology* 2006;28:29–42. [PubMed: 16448974]
- Levine A, Miller E, Becker J, Selnes O, Cohen B. Normative data for determining significance of test–retest differences on eight common neuropsychological instruments. *The Clinical Neuropsychologist* 2004;18:373–384. [PubMed: 15739809]
- 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;43:1551–1558. [PubMed: 12460258]
- Miller EN, Selnes OA, McArthur JC, Satz P, Becker JT, Cohen BA, et al. Neuropsychological performance in HIV-1 infected homosexual men: The Multicenter AIDS Cohort Study (MACS). *Neurology* 1990;40:197–203. [PubMed: 2405289]
- Hasin DS, Trautman KD, Miele GM, Samet S, Smith M, Endicott J. Psychiatric Research Interview for Substance and Mental Disorders (PRISM): Reliability for substance abusers. *American Journal of Psychiatry* 1996;153:1195–1201. [PubMed: 8780425]
- Rippeth J, Heaton R, Carey C, Marcotte T, Moore D, Gonzalez R, et al. Methamphetamine dependence increases risk of neuropsychological impairment in HIV infected persons. *Journal of the International Neuropsychological Society* 2004;10:1–14. [PubMed: 14751002]
- Salinsky M, Storzach D, Dodrill C, Binder L. Test–retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12–16 week period. *Journal of the International Neuropsychological Society* 2001;7:597–605. [PubMed: 11459111]
- Sawrie S, Chelune G, Naugle R, Luders H. Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. *Journal of the International Neuropsychological Society* 1996;2:556–564. [PubMed: 9375160]
- Sherman E, Slick DJ, Connolly MB, Steinbok P, Martin R, Strauss E, et al. Reexamining the effects of epilepsy surgery on IQ in Children: Use of regression-based change scores. *Journal of the International Neuropsychological Society* 2003;9(6):879–886. [PubMed: 14632247]
- Smith, A. Symbol Digit Modalities Test. Los Angeles: Western Psychological Services; 1991.
- Stroop JR. Studies of interference in serial verbal reaction. *Journal of Experimental Psychology* 1935;8:643–662.
- 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;5:357–369. [PubMed: 10349298]
- Wechsler, D. WAIS–R manual. New York: The Psychological Corporation; 1981.

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

TABLE 1

MACS nonclinical (HIV-negative) sample characteristics

Measure	N	Age ^{a,b}	Education ^{a,b}	Interval ^{a,c}
Grooved Pegboard	605	39.5 (8.7)	16.4 (2.3)	218 (95)
Trail-Making Test	344	41.4 (9.3)	16.3 (2.3)	235 (127)
Stroop (Kaplan)	439	42.8 (8.9)	16.5 (2.3)	256 (133)
Digit Span	674	37.6 (7.6)	16.4 (2.3)	221 (98)
SDMT	465	38.1 (7.8)	16.3 (2.3)	192 (53)
COWAT	145	38.8 (7.9)	16.4 (2.3)	191 (38)

Note. SDMT = Symbol Digit Modalities Test. COWAT = Controlled Oral Word Association Test.

^a Means; standard deviations in parentheses.

^b In years.

^c Days between baseline assessment and first follow-up assessment.

TABLE 2

MACS HIV-seropositive sample characteristics

Characteristic	Mean	SD	% of participants
Age ^a	36.5	6.7	
Education ^a	15.5	2.4	
Estimated intelligence ^b	107	10	
BDI ^c	9	7.5	
CD4+ T-cells	473	277.5	
Ethnicity			
African American			10.4
Caucasian			81.6
Hispanic			6.6
Asian or Native American			1.4
Multiracial			3.3
Female			0
CDC AIDS diagnosis			57
Current substance dependence ^d			0

Note. MACS = Multicenter AIDS Cohort Study. CDC = Centers for Disease Control.

^aIn years.

^bEstimated IQ based on the American version of the Nelson Adult Reading Test (Grober & Sliwinski, 1991).

^cBeck Depression Inventory (Beck, 1987).

^dMet criteria within the last month per DSM-IV (American Psychiatric Association, 1994).

TABLE 3

Clinical (HIV+) MACS cohort sample characteristics by test

Measure	CDC group	N	Age ^a	Education ^a	Interval ^b
Grooved Pegboard	<200	153	37.1 (6.4)	15.5 (2.5)	254 (134)
	200–499	346	37 (6.8)	15.8 (2.3)	234 (113)
	≥500	282	35.8 (6.7)	15.7 (2.3)	235 (117)
Trail Making Test	<200	204	38.3 (6.9)	15.5 (2.4)	239 (117)
	200–499	415	37.3 (7)	15.8 (2.3)	239 (117)
	≥500	331	36.2 (6.7)	15.8 (2.2)	231 (110)
Stroop (Kaplan)	<200	147	39.5 (6.4)	15.3 (2.4)	245 (116)
	200–499	178	40 (6.8)	15.8 (2.7)	246 (125)
	≥500	121	39.4 (7.3)	15.8 (2.5)	245 (126)
Digit Span	<200	111	36.3 (5.8)	15.6 (2.6)	265 (144)
	200–499	280	36.7 (6.8)	15.8 (2.3)	229 (106)
	≥500	232	35.2 (6.6)	15.7 (2.3)	239 (122)
SDMT	<200	183	37.8 (6.9)	15.6 (2.4)	244 (126)
	200–499	403	36.7 (7)	15.7 (2.3)	231 (109)
	≥500	364	35.7 (6.7)	15.7 (2.3)	228 (107)
COWAT	<200	111	36.3 (5.8)	15.6 (2.6)	258 (139)
	200–499	279	36.6 (6.8)	15.8 (2.3)	227 (104)
	≥500	233	35.3 (6.6)	15.7 (2.3)	238 (122)

Note. MACS = Multicenter AIDS Cohort Study. CDC = Centers for Disease Control. SDMT = Symbol Digit Modalities Test. COWAT = Controlled Oral Word Association Test.

^aIn years.

^bIn days.

TABLE 4

NIDA sample demographic characteristics

Characteristic	Mean	SD	% of participants
Age ^a	41.7	7.0	
Education ^a	13	2.1	
Estimated intelligence ^b	105	9.1	
BDI-2 ^c	13.6	9.2	
CD4 (self-report)	461	307	
Ethnicity			
African American			68
Caucasian			13.5
Hispanic			11
Asian or Native American			4.5
Multiracial			3
Female			16
CDC AIDS diagnosis			65
Current substance dependence ^d			32

Note. NIDA = National Institute of Drug Abuse. CDC = Centers for Disease Control.

^a In years.

^b Estimated IQ based on the American version of the Nelson Adult Reading Test (Grober and Sliwinski, 1991).

^c Beck Depression Inventory, Second Edition (Beck, Steer, & Brown, 1996).

^d Met criteria for alcohol or drug abuse/dependence within the last month per DSM-IV (American Psychiatric Association, 1994), based on Psychiatric Research Interview for Substance and Mental Disorders (PRISM; Hasin et al., 1996) interview at baseline visit.

TABLE 5

Test scores for both samples at baseline and follow-up

Measure	NIDA			MACS		
	Baseline score ^a	Retest score ^a	p-value	Baseline score ^a	Retest score ^a	p-value
Trail Making Test: Form A ^b	34.1 (13.8)	31.3 (9.3)	<.001	25.4 (8.5)	23.7 (7.6)	<.001
Trail Making Test: Form B ^b	91.9 (50.3)	85.5 (37.6)	<.001	55.7 (20.3)	52.8 (19.4)	<.001
Grooved Pegboard Dominant hand ^b	78.5 (22.4)	80 (21.5)	<.001	63.2 (8.2)	61.9 (9.3)	<.001
Grooved Pegboard Nondominant hand ^b	87 (26.8)	91.3 (37.3)	<.001	68.9 (10.8)	67.2 (12)	<.001
COWAT ^c	37.1 (11.9)	37.8 (12.5)	<.001	45.3 (11.5)	46.4 (11.5)	<.001
Digit Span ^d	15 (3.9)	14.6 (3.9)	<.001	17.1 (4.4)	17.3 (4.2)	.01

Note. NIDA = National Institute of Drug Abuse. MACS = Multicenter AIDS Cohort Study. COWAT = Controlled Oral Word Association Test.

^aMeans, with standard deviations in parentheses.

^bTime in seconds.

^cTotal words across three trials (FAS).

^dTotal digits (backward & forward).

TABLE 6

Rates of change among the MACS sample

	CDC Group	Regression-based classification			RCI-based classification		
		Decline	Stable	Improve	Decline	Stable	Improve
COWAT	0-200	8	90.9	1.1	10.2	87.5	2.3
	201-499	9.6	86.6	3.8	10.3	86.2	3.4
	500+	7.4	88	4.6	10.1	84.8	5.1
	Total %	8.5	87.8	3.7	10.2	85.9	3.9
Digit span	0-200	4.1	90.5	5.4	8.1	87.8	4.1
	201-499	6.6	87.7	5.7	9.7	82.8	7.5
	500+	5.1	89.1	5.8	8.3	84.6	7.1
	Total %	5.7	88.6	5.7	9	84.2	6.8
Trail Making Test: Form A	0-200	12.3	84.2	3.5	4.4	86.2	9.4
	201-499	9.9	86.8	3.3	1.5	84.8	13.7
	500+	6	90.5	3.5	0.6	85.7	13.7
	Total %	9.1	87.5	3.4	1.9	85.4	12.7
Trail Making Test: Form B	0-200	14.4	82.2	3.5	3.5	85.1	11.4
	201-499	10.7	87.3	2.0	3.6	87.8	8.6
	500+	1.9	87.8	1.9	2.3	91.3	6.4
	Total %	11.3	86.3	2.3	3.1	88.4	8.5
Grooved Pegboard: dominant hand	0-200	11.1	84.4	4.4	7.8	82.2	10
	201-499	7.1	89.6	3.3	1.8	87	11.2
	500+	7.3	90	2.7	2.3	85.8	11.9
	Total %	7.8	88.9	3.3	2.3	85.8	11.9
Grooved Pegboard: nondominant hand	0-200	12.2	84.4	3.3	5.6	78.9	15.6
	201-499	7.4	87.4	5.2	1.5	81.8	16.7
	500+	10.5	86.8	2.7	1.8	82.6	15.5
	Total %	9.3	86.7	4	2.2	81.7	16.1

Note. MACS = Multicenter AIDS Cohort Study. RCI = reliable change indexes. CDC = Centers for Disease Control. COWAT = Controlled Oral Word Association Test.

TABLE 7

Rates of change among NIDA participants

	CDC group	Regression-based classification			RCI-based classification		
		Decline	Stable	Improve	Decline	Stable	Improve
COWAT	0–200	19.4	80.6	0	0	90.3	9.7
	201–499	16.4	80.8	2.7	0	94.5	5.5
	500+	22.4	77.6	0	1.7	98.3	0
	Total %	19.1	79.7	1.2	0.6	95.1	4.3
Digit Span	0–200	9.7	90.3	0	0	100	0
	201–499	12.3	83.6	4.1	4.1	95.9	0
	500+	13.8	82.8	3.4	5.2	94.8	0
	Total %	12.3	84.6	3.1	3.7	96.3	0
Trail Making Test: Form A(N = 168)	0–200	26	71	3	6.5	90.3	3.2
	201–499	21.6	75.7	2.7	8.1	85.1	6.8
	500+	20.5	76	3.5	6.9	91.4	1.7
	Total %	22.1	74.8	3.1	7.4	88.3	4.3
Trail Making Test: Form B	0–200	26.7	66.7	6.6	16.7	70	13.3
	201–499	33.7	62.2	4.1	8.1	85.1	6.8
	500+	40.7	50.8	8.5	11.9	74.6	13.5
	Total %	35	58.9	6.1	11	78.6	10.4
Grooved Pegboard: dominant hand	0–200	48.4	51.6	0	3.2	96.8	0
	201–499	39.2	58.1	2.7	12.2	83.8	4.1
	500+	37.3	62.7	0	18.6	81.4	0
	Total %	40.2	58.6	1.2	12.8	85.4	1.8
Grooved Pegboard: nondominant hand	0–200	45.2	51.6	3.2	9.7	87.1	3.2
	201–499	50	48.6	1.4	21.6	74.3	4.1
	500+	31	62.1	6.9	10.3	84.5	5.2
	Total %	42.3	54	3.7	15.3	80.4	4.3

Note. NIDA = National Institute of Drug Abuse. RCI = reliable change indexes. CDC = Centers for Disease Control. COWAT = Controlled Oral Word Association Test.