An Evaluation of Methods for Measuring Cognitive Change in Older Adults

by

Robert John Frerichs
B.A., University of Saskatchewan, 1994
M.Sc., University of Victoria, 1998

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Psychology

We accept this dissertation as conforming
to the required standard

Dr. H. Tuokko, Supervisor (Department of Psychology)

Dr. R. E. Graves, Departmental Member (Department of Psychology)

Dr. C. Mateer, Departmental Member (Department of Psychology)

Dr. L. Gamroth, Outside Member (School of Nursing)

Dr. G. E. Smith, External Examiner (Department of Psychiatry and Psychology, Rochester Mayo Clinic)

University of Victoria

Supervisor: Dr. Holly A. Tuokko

Abstract

Serial neuropsychological assessment of older adults requires well-researched statistical methods to guide clinicians in determining the significance of test score changes at the individual level. This study examined the standard deviation (SD) method, various reliable change indices (RCIs), and three standardized regression-based (SRB) methods in older adults who participated in one or more waves of the Canadian Study of Health and Aging (CSHA). Changes in test scores were examined in cognitively healthy older adults over a short test-retest interval of a few months and a longer interval spanning approximately 5 years. Test score changes were also compared to clinically significant indices including change in diagnostic status, subjective report of loss, informants' ratings of loss, and clinicians' rating of loss. The findings indicated that practice effects were not a prominent feature of older adults' performance. Mean decline was shown on neuropsychological tests of memory and psychomotor speed over a test-retest interval of approximately 5 years. At the individual level, normal variability in the test performance of cognitively healthy adults could be accurately classified using several methods over a short interval but only select methods over the longer interval. Two RCIs and three SRB methods were relatively accurate in classifying change among persons who remained cognitively intact and in those who had progressed to a dementia by follow-up 5 years later. A combination of memory measures and these change score methods resulted in diagnostic classification accuracy of approximately 89% in this sample. Diagnostic accuracy was also significantly associated with the sum of reliable test score changes using different change score methods. Reliable deterioration was moderately associated with clinicians' and informants' ratings of cognitive

loss and weakly associated with subjective ratings of memory loss. These findings have implications for clinicians who seek to determine the meaning of neuropsychological test-retest score changes.

Examiners:

Dr. H. Tuokko, Supervisor (Department of Psychology)

Dr. R. E. Graves, Departmental Member (Department of Psychology)

Dr. C. Mateer, Departmental Member (Department of Psychology)

Dr. L. Gamroth, Outside Member (School of Nursing)

Dr. G. E. Smith, External Examiner (Department of Psychiatry and Psychology, Rochester Mayo Clinic)

Table of Contents

List of Tables

Acknowledgments

Among the many individuals who deserve mention for their contribution, I would like to first acknowledge all of the Canadian Study of Health and Aging participants who were kind enough to volunteer their time and energy to take part in this study. The project itself would not be possible without the funding provided by the Alzheimer's Society of British Columbia, to whom I am deeply indebted.

I am grateful to all members of my supervisory committee for their support and valuable feedback. However, I extend a special thanks to Dr. Tuokko for her assistance, guidance, and unwaivering support throughout my training. As my mentor, she has been invaluable in my development as a clinician and a researcher.

Dedication

To my wife,
for her continued love and support.

Geriatric neuropsychology is one of the fastest growing areas within the field of clinical neuropsychology. There are a variety of reasons for this trend. First, cognitive functioning is a common concern in the later stages of life. As people grow older, many experience normal age-related changes in cognitive, motor, and sensory functioning. A greater mental health concern stems from the fact that older adults also have relatively high rates of medication use and a heightened risk for medical conditions such as Alzheimer Disease that may cause significant cognitive impairment. Second, changes in the demographic structure of society and the "graying" of the nation (Statistics Canada, 1991) have resulted in an increased demand for neuropsychological services among the population of adults over age 65. The need for these services is likely to extend well into the future. Third, neuropsychological research continues to contribute to our understanding of normal and pathological aging processes. Several measures have been developed or adapted for use with older adults in clinical and research settings (Libon et al., 1996; Mattis, 1988; Morris, Heyman, & Mohs, 1989; Tuokko, Hadjistavropoulos, Miller, Horton, & Beattie, 1995). There has also been an increase in the availability of normative information specific to older adults within the last decade (Ivnik, Malec, Smith, Tangalos, & Petersen, 1996; Ivnik et al., 1992; Ivnik et al., 1997; Lucas, Ivnik, Smith, Bohac, Tangalos, Graff-Radford et al., 1998; Lucas, Ivnik, Smith, Bohac, Tangalos, Kokmen et al., 1998; Paolo, Troster, & Ryan, 1997; Ryan, Paolo, & Brungardt, 1990; Tombaugh & Hubley, 1997; Tuokko & Hadjistavropoulos, 1998).

The popularity and demand for geriatric neuropsychological assessment is driven by the many purposes that the information can serve. An evaluation of an individual's pattern of cognitive strengths and weaknesses may inform decisions about diagnosis or

the management and planning of a client's care (Lezak, 1995). Another particularly important role of neuropsychological assessment is in measuring progress or decline in cognitive functioning over time. In some instances, the measurement of change itself may contribute to establishing a particular diagnosis. Dementia and delirium are examples of disorders that require evidence of cognitive decline over time or fluctuations in cognition, respectively, to be clinically diagnosed (American Psychiatric Association, 1994). A decline in cognitive functioning may also have prognostic implications if it heralds the onset of a dementing illness (Crystal et al., 1996; Howieson et al., 1997; Jacobs et al., 1995; Masur, Sliwinski, Lipton, Blau, & Crystal, 1994; Rubin et al., 1998). But not all decreases in cognitive functioning are diagnostic of dementia or predictive of its subsequent development. The critical task facing clinicians and researchers is how to distinguish normal cognitive change from change that is clinically relevant for any given individual.

Identifying cognitive change from a single assessment can be difficult but most dementia assessments are fashioned in this manner. Imagine a situation in which a clinical neuropsychologist finds a pattern of weakness in multiple cognitive domains. This pattern might reasonably be interpreted as evidence of significant cognitive decline in a person who was presumed to be cognitively intact at an earlier stage in life. The inherent difficulty in drawing such a conclusion from a single "snapshot" is that it does not rule out competing explanations for the poor performance and therefore may lead to a classification error. It is equally viable, for instance, that this individual had always exhibited poor memory and language skills (i.e., a level of performance in the lower end of the normal distribution). An increased potential for misclassification also exists for

individuals in the early stages of a dementing process. Individuals who have high premorbid IQs and/or high levels of education may evidence substantial decline as part of the initial stages of Alzheimer Disease but still score well within the average range on a cognitive measure. Clinical neuropsychologists who interpret the "normal" test scores but fail to detect deterioration in cognition may be doing a serious disservice to their clients. These examples illustrate the limitation of using singular assessments to derive information about changes in functioning.

To overcome these difficulties, indirect methods are often employed from which cognitive change may be inferred (Lezak, 1995). Two common methods are to: 1) use collateral sources to provide information about prior functioning and subsequent changes, or 2) make a comparison of current functioning to an estimate of premorbid functioning (Graves, 2000; Graves, Carswell, & Snow, 1999; Paolo & Ryan, 1992; Schinka & Vanderploeg, 2000). Collateral information, though important, may not be available in all cases and in some instances it may be biased to either overestimate or underestimate actual changes in cognition. An estimate of premorbid functioning is an important part of any dementia assessment but by its very nature lacks the precision of formal assessment. So while the methods described above are immensely useful in practice, it is generally preferable to directly and objectively evaluate change in a person's cognitive functioning by taking measurements on two or more occasions.

The measurement of change over time using repeated assessments is of increasing importance in clinical neuropsychology. The popularity of follow-up assessments has particularly grown among certain populations. With older adults, serial neuropsychological assessment has been useful as a means to monitor disease

progression (Brull, Wertheimer, & Haller, 1979; McCleary, Dick, Buckwalter, Henderson, & Rodman Shankle, 1996; Morris et al., 1989; Morris et al., 1993; Storandt, Botwinick, & Danziger, 1986; Taylor, 1998). Multiple evaluations are also believed to improve the accuracy of dementia diagnosis and the early detection of cognitive decline relative to the single assessment (Cummings & Benson, 1983; Flicker, Ferris, & Reisberg, 1993; Mitrushina & Satz, 1991). In addition, the measurement of neuropsychological change has played a key role in assessing recovery following brain injury (e.g., Dikmen, Machamer, Temkin, & McLean, 1990; Hinton-Bayre, Geffen, Geffen, McFarland, & Friis, 1999; Iverson, 1999; Wilson, Watson, Baddeley, Emslie, & Evans, 2000), the impact of a medical treatment (e.g., Cahn et al., 1998; Helmstaedter, Gleibner, Zentner, & Elger, 1998; Hermann et al., 1996; Kneebone, Andrew, Baker, & Knight, 1998; Purdon et al., 2000; Weinstein et al., 1999), and the effectiveness of rehabilitation programs (e.g., Chen, Thomas, Glueckauf, & Bracy, 1997; Scherzer, 1986; Sohlberg & Mateer, 1989a; Sohlberg & Mateer, 1989b).

Several questions need to be addressed if clinicians are to have an informed understanding about how to interpret cognitive change in the context of repeated assessments. For example, how much change is normal at retest? How much change is abnormal? Is abnormal change diagnostic of conditions such as dementia? If so, are particular methods of measuring change more appropriate for use with specific populations or specific measures than others? What effect does the test-retest interval length have on these methods? Researchers who are interested in answering these questions need to consider some of the following issues associated with serial neuropsychological assessments.

One primary concern is that measures sensitive to cognitive deficits and underlying brain dysfunction may not be useful in change measurement. Some neuropsychological measures exhibit floor or ceiling effects that dramatically limit the degree to which an individual's score might decline or improve on repeated assessment. Measures that are not sensitive to a full range of performance within a cognitive domain may be inappropriate for change measurement. Another important consideration is that neuropsychological measures, on the whole, have not been validated for the purpose of detecting change. The sensitivity of most neuropsychological tests to cognitive deficits has been established through innumerable investigations showing statistically significant group differences but the measurement of change in serial assessments is focused on detecting clinically significant differences within the individual. Group-level statistical comparisons of change data (e.g., matched t-tests or repeated measures ANOVA) examine mean pretest-posttest changes and tend to obscure the variability at the individual-level of analysis that is of primary interest to the clinician (Jacobson, Roberts, Berns, & McGlinchey, 1999; Phillips & McGlone, 1995).

An example may illustrate the importance of empirically-validating an instrument for measuring change. The Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) is an 11-item measure that is moderately accurate in discriminating between groups of persons with and without dementia (Tombaugh & McIntyre, 1992). It is frequently employed to screen for dementia and cognitive impairment at the individual-level. Due to its ease of administration and brief length, the MMSE is also commonly used by mental health professionals to monitor cognitive status over time. Recent evidence, however, suggests that the MMSE may be of limited value in tracking

cognitive change in persons with Alzheimer Disease who are followed up for less than 3 years (Clark et al., 1999). This is largely because the amount of measurement error associated with the MMSE is nearly equal to the average annual rate of change. This fact makes it difficult for clinicians to distinguish between change due to random error and change that is meaningful. Clarke et al.'s (1999) findings pertaining to the MMSE are relevant to all neuropsychological instruments that might also be used to measure change. There is a need for neuropsychologists to establish, rather than assume, the sensitivity of their measures to intraindividual changes.

Information regarding normal variability in test performance over time is currently lacking for most neuropsychological measures. This is particularly true for older adults for whom the effects of multiple assessments and age-related cognitive decline are less than well understood. Investigations involving serial evaluations of older adults reveal overall mean group changes in cognitive functioning over time (e.g., Frank, Wiederholt, Kritz-Silverstein, Salmon, & Barrett-Connor, 1996; Mitrushina & Satz, 1991; Schaie, 1996) but these studies do not provide information about the degree of variability at the individual-level. The same holds true for most studies that yield an estimate of the test-retest reliability for specific neuropsychological measures. Changes in relative ranking and group mean changes with age do not adequately capture individual variability across the lifespan. Some older individuals change very little over time whereas others change dramatically. These changes may conform to a linear pattern or they may comprise discrete periods of stability mixed with marked increases and declines. Accumulating evidence suggests that variability in cognitive test performance increases with age on some tasks but not others (Christensen et al., 1999; Hertzog, Dixon,

& Hultsch, 1992; Shammi, Bosman, & Stuss, 1998). Data pertaining to the normal individual variability that may be expected as part of serial neuropsychological evaluations are beginning to appear in the literature (e.g., Ivnik et al., 1999) but more research is required to inform clinical decisions about the significance of observed cognitive change for a given person.

Another major limitation to the serial assessment approach is that there is no consensus regarding how change in performance should be measured at the individual case level. Methodologies for studying individual change emerged nearly 50 years ago (e.g., Harris, 1963; Lord, 1957, 1958; McNemar, 1958; Payne & Jones, 1957) and refined methods continue to appear in the literature (e.g., Crawford & Howell, 1998; Hageman & Arrindell, 1999b; Hsu, 1989; Jacobson & Truax, 1991). The best methods for measuring change have been the subject of frequent debate (Cronbach & Furby, 1970; Maassen, 2000b, 2001; Rogosa, 1988; Rogosa, Brandt, & Zimowski, 1982). This, in turn, has generated confusion in the literature that has yet to be resolved (Speer, 1999).

In short, clinical neuropsychologists who work with elderly clients must address many important issues if useful information is to be derived from serial assessments. There is a clear need to examine change measurement methods, validate neuropsychological measures for studying change, and collect data regarding the amount of change that is normal and abnormal in older adults over varying intervals. Until these issues are addressed, obstacles to the assessment and interpretation of cognitive change in late life will persist.

Confounds in the assessment of change

A neuropsychological test with established validity for the purpose of measuring change should be sensitive to true changes in cognition rather than changes that arise due to confounding factors. Many confounding factors can complicate the interpretation of individual change in test-retest designs (Chelune, 1998). Random errors may occur as a result of the unreliability of a measure and statistical effects such as regression to the mean. Bias, referring to any systematic distortion in measurement, may also make the interpretation of change difficult. A common bias in serial assessment is the practice effect. What follows is a review of some of the important errors and biases inherent in change measurement and their interaction with normal age-related cognitive decline.

Reliability. Reliability is a central concept in measurement theory that broadly refers to the consistency, stability, or repeatability of test scores from parallel measures. Classical test theory (CTT; Gulliksen, 1950; Lord & Novick, 1968) posits that an observed score is a combination of an individual's true score and an unspecified amount of measurement error. The amount of measurement error is a function of the measure's reliability. Unreliable measures yield virtually meaningless test scores with large measurement errors whereas perfectly reliable measures provide an exact indication of a person's true score. In the measurement of change, the difference between scores obtained on two separate occasions only reflects "true" change if the measure has perfect reliability; that is, the capacity to provide an unbiased estimate of the person's score that is free from all errors of measurement (Lord, 1963). According to CCT, all neuropsychological measures are associated with some degree of unreliability and minor random fluctuations are to be expected as a consequence. Therefore, one can not expect an individual to obtain the same results on a neuropsychological test when it is completed

a second time, even when there has been no real change in the client's cognitive abilities. There is no relation between age and measurement error (at least in theory) since errors due to unreliability are random. As discussed later, the methods used to assess change in test scores differ in the extent to which they account for measurement errors due to unreliability.

A variety of formulae exist to estimate the reliability of a measure, each of which yields a different reliability coefficient (see Anastasi, 1988; Lord & Novick, 1968; Nunnally, 1978; Stanley, 1971). Error variance comes from several sources and different reliability coefficients may be calculated to reflect the degree of agreement among persons (inter-rater reliability), agreement among the test items (internal consistency), or the temporal stability of the measure (test-retest reliability). CTT only allows control of one source of error at a time and, unfortunately, there is considerable debate regarding the most appropriate choice of reliability coefficient (Streiner & Norman, 1995). The test-retest method is common but has been viewed as the least appropriate method for estimating the reliability of a measure (Anastasi, 1988; Nunnally, 1978). Test-retest reliability coefficients vary considerably depending upon the population studied, the sample size, and the length of the test-retest interval. Test-retest intervals that are too brief may be unduly influenced by memory for responses to specific test items. With longer test-retest intervals, the true score of the individual is more likely to change as a result of aging, a disease process, or a specific treatment thereby hindering any attempt to estimate the temporal stability of a measure. Internal consistency coefficients are generally viewed as more stable estimators of a measure's reliability (Nunnally, 1978) when a single test score is the focus of study. This approach, however, may be less

appropriate than using the test-retest correlation coefficient when the focus is on two test scores, as with serial assessments.

To estimate the amount of error inherent in test scores, the reliability of the measure must be expressed in terms of a standard error value associated with that measure. Various standard error terms exist; most of which have been frequently misinterpreted and misused (Brophy, 1986; Charter, 1996; Dudek, 1979; Glutting, McDermott, & Stanley, 1987). Using Lord and Novick's (1968) conventions, these error terms include the standard error of measurement (SEM), the standard error of estimate (SEE), and the standard error of prediction (SEP). All standard error terms are a function of reliability and therefore, vary depending upon the specific reliability coefficient that one chooses to use. The SEM is defined as $SD_X (1 - r_{xx})^{1/2}$ where SD_X is the standard deviation of the measure and r_{xx} is the reliability coefficient. It is an index of the dispersion of an obtained score about an unknown true score. Since one rarely has knowledge of an individual's true score, it is usually inappropriate for clinical use (though often employed). The SEE is similar to the SEM but refers to the distribution of true scores if an obtained score is held constant. The SEE is smaller than the SEM and is defined as $SD_X (r_{xx} (1 - r_{xx}^2))^{1/2}$. The SEE, rather than the SEM, is the appropriate error term when one wishes to form confidence intervals around an estimated true score (e.g., to determine the interval that would bound a person's true IQ). It should be noted that this definition of SEE differs from that used in regression analyses where SEE = $(SS_{residuals} / N - k - 1)^{1/2}$. Finally, the SEP, the largest error term, is defined as $SD_X (1 - r_{xx}^2)^{1/2}$. The SEP refers to the distribution of observed test scores if the observed test scores from a parallel version of the measure are held constant. Accordingly, the SEP

has been identified as the appropriate error term to use in test-retest situations (Brophy, 1986; Charter, 1996; Dudek, 1979). As discussed later, the standard errors terms and their appropriate use are important considerations in the determination of change in neuropsychological test performance over time.

Regression to the mean. Regression to the mean is a statistical phenomenon that is closely associated with reliability. Regression to the mean refers to the tendency for baseline scores, particularly those at either the high or low end of a distribution, to move toward the mean upon retesting (Nesselroade, Stigler, & Baltes, 1980). It is predominantly linked to test-retest designs in which the reliability of the measuring device is less than perfect (i.e., $r < 1.0$). Regression to the mean occurs because extreme scores are either comprised of an unusually large proportion of error or they arise from a relatively rare combination of antecedent events (Nesselroade et al., 1980). In either case, the factors leading to the production of extreme scores at time 1 are unlikely to be maintained at follow-up meaning that the time 2 score is more likely to be closer to the overall mean.

Regression to the mean is often discussed in relation to measuring change, but it is not well understood by researchers and clinicians (Gottman & Rushe, 1993). Rogosa (1988), for example, challenged the statistical myth that these effects are ubiquitous in longitudinal data. Regression to the mean is a statistical feature of any linear prediction rule utilizing a "least squares" model, but it may be avoided in some instances if it is defined in a metric other than standard deviation units. When redefined in this manner, Rogosa suggested that regression to the mean only occurs when the correlation between true change and initial true score is negative. Unfortunately the correlation between

observed baseline score and observed change provides a poor estimation of these population parameters (Rogosa et al., 1982) making it difficult to determine when in fact regression to the mean effects are active. At best, it may be stated that regression to the mean is a potential complicating factor in the assessment of change using test-retest data. However, there is uncertainty and confusion regarding its pervasiveness and its actual impact on the interpretation of clinical data (McGlinchey & Jacobson, 1999; Speer, 1999).

Practice effects. Practice effects refer to observed improvements in test performance that are solely due to repeated assessment with the same instrument; they are typically operationalized in terms of overall group mean change between two testing occasions (McCaffrey & Westervelt, 1995). The effects of practice vary as a function of the measuring device, the retest interval, the number of previous exposures, and characteristics of the individual (e.g., age, history of head injury, ability to learn). With regard to neuropsychological measures, practice effects are generally greatest with timed tests, those requiring an infrequently practiced response, or those having a single, easily conceptualized solution (Dodrill & Troupin, 1975; Lezak, 1995). Information about practice effects is not available for most neuropsychological measures (for some exceptions, see Basso, Bornstein, & Lang, 1999; Matarazzo, Carmody, & Jacobs, 1980; McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992; Shatz, 1981) but it is generally believed that the influence of practice is minimized as the test-retest interval increases. The amount of time that must pass before practice effects become negligible is unknown though there is some indication that the effect may operate for up to six years (Zelinski & Burnight, 1997). Research findings suggest that practice effects tend to disappear after

the second testing session (Ivnik et al., 1999; Theisen, Rapport, Axelrod, & Brines, 1998).

The relation between age and practice effects warrants special consideration (Horton, 1992; McCaffrey & Westervelt, 1995). In contrast to the positive effects of practice, normal aging is associated with an overall drop in test performance across a range of cognitive domains (Albert, 1994; Flicker, Ferris, Crook, Bartus, & Reisberg, 1985; Korten et al., 1997). In combination these opposing effects work to simply cancel each another out. Age-associated cognitive decline may also restrict an older person's ability to benefit from prior exposure to a test. Research supports the notion that practice effects decrease with advancing age, especially in persons over age 75 (Mitrushina & Satz, 1991; Ryan, Paolo, & Brungardt, 1992). The interaction between age and practice effects is important because it may inform the interpretation of change, particularly if the absence of practice effects has diagnostic value (Lezak, 1995; McCaffrey, Ortega, & Haase, 1993). For example, consider two cognitively healthy persons. One individual is age 60 and another is age 80. Both obtain age-corrected retest scores that are the same as their age-corrected baseline scores. Though the lack of apparent change in test scores suggests that both remained cognitively normal, an informed clinician might interpret the absence of a practice effect in the 60-year-old individual as evidence of a cognitive deterioration. The reason is that practice effects are to be expected at age 60 on this hypothetical measure. The stability of the 80-year-old person's score, in contrast, is attributable to both aging and practice (and not a dementing process). The point to be made is that practice effects and their interaction with other variables are important, but

often overlooked, considerations in serial neuropsychological assessment and the determination of meaningful change.

The measurement of change

Over the past decade, various statistical methods have been proposed to minimize or account for the errors and biases inherent in multiple assessments. The following review will focus on those methods designed to measure change over two occasions (i.e., test-retest designs). Change, arguably, is most meaningfully examined through the collection of multi-wave data employing more than two measurements (Rogosa, 1988; Rogosa et al., 1982; Speer, 1999; Speer & Greenbaum, 1995), but there are instances when this is neither feasible nor appropriate (Hageman & Arrindell, 1999a). The test-retest design remains common in the neuropsychological literature and pertinent to clinical practice (particularly with older adults). Therefore, it is worthwhile to consider the variety of methods for studying change using two-wave data.

Simple difference method. The difference in observed scores between pretest and posttest is the most obvious and simple measure of change. It is also the most maligned. Difference scores have been frequently criticized as poor indicators of change due to low reliability and their tendency to correlate negatively with initial status (Cronbach & Furby, 1970; Linn & Slinde, 1977; Lord, 1963). Under circumstances where the standard deviation and reliability of the measurement instruments do not change over time, it has been shown that the reliability of the difference score tends to decrease as the pretest-posttest correlation increases. The implication is that the use of neuropsychological measures with high test-retest reliability may not yield reliable difference scores. A second criticism is that persons with low (or high) scores on a certain measure are more

likely to exhibit large (or small) difference scores. This relation would appear to "give an advantage to persons with certain values of the pretest score" (Linn & Slinde, 1977, p. 125) making the use of difference scores untenable.

Rogosa (1988; 1982) has challenged both criticisms and defended the use of the difference score as an unbiased estimate of true change. He argued that difference scores are not intrinsically unreliable; they are only unreliable if there is little variability in change rates across persons. The reliability of a difference score is quite respectable so long as there are individual differences in true change within the population of interest. Furthermore, Rogosa viewed the negative correlation between initial status and change ($r_{X1\cdot D}$) as an irrelevant artifact arising from errors of measurement. The correlation between an observed pretest score and observed change (both of which are subject to measurement error) provides an inadequate and biased estimate of the population correlation between initial true score and true score change (i.e., the correlation of real interest). He did not view the negative bias of $r_{X1\cdot D}$ an obstacle to using the difference score as a measure of individual change.

Several measures of change that will be discussed are linear transformations of the difference score involving a standard error term. For the difference score to be used as an indicator of "significant" or diagnostic change requires a cut-off point. According to the following formula, a difference score (D) greater than a specified cutoff value (CV) is considered to reflect significant deterioration whereas change failing to meet this criterion is not.

$$D = X_2 - X_1 > CV \qquad \text{(Equation 1)}$$

X_1 = observed pretest score and X_2 = observed posttest score.

Matarazzo, Carmody, and Jacobs' (1980) rule of thumb exemplifies this approach. These authors suggested that a change of at least 15 points in IQ must be evident before interpreting a change as "potentially" clinically important. One of the main drawbacks to this approach is that cutoff scores may be arbitrarily chosen or selected on the basis of idiosyncratic criteria. As such, they do not take account of the magnitude of measurement error or the presence of practice effects. In practice, cutoff scores that are empirically-informed are sample specific. That is, they may vary as a function of the sample from which they are derived and may not generalize well to clinical settings.

Standard deviation method. A second approach to defining change in cognitive functioning is the standard deviation (SD) method in which a client is considered to have deteriorated if his/her difference score is more than 1 SD of the group pretest score on a certain measure. The formula for significant change using this method is as follows:

$$C = X_2 - X_1 / SD_1 \qquad \text{(Equation 2)}$$

SD_1 is the standard deviation of the pretest scores and X_1 and X_2 are as previously defined. For measures in which a higher score reflects improved performance, $C > 1.0$ is indicative of significant improvement and $C < -1.0$ is indicative of significant deterioration. The opposite pattern holds for measures in which a lower score reflects better performance.

The use of 1 SD as the criterion for cut-off appears to be arbitrary since it is not clearly informed by any sound psychometric consideration, such as establishing a desired specificity. In practice, the SD method has been used to assess neuropsychological change following temporal lobectomy and cardiac surgery (Hermann & Wyler, 1988;

Mahanna et al., 1996; Phillips & McGlone, 1995; Shaw et al., 1986). It has also been used to classify cognitive change in persons with and without dementia (Bieliauskas, Fastenau, Lacy, & Roper, 1997). Though the method is simple, there is little consistency in how the approach is applied. Some studies treat a significant decline on a single test as evidence of change whereas others operationalize change as a decline of 1 SD on 20% of all measures administered. The SD method for detecting change in test-retest scores has been criticized for its failure to account for measurement errors in the observed scores and the effects of practice.

Reliable change indices. Despite cautions about its appropriate use (Brophy, 1986; Charter, 1996; Dudek, 1979), the SEM has been advocated as an acceptable method for estimating the significance of test-retest changes in the individual (Edwards, Yarvis, Mueller, Zingale, & Wagman, 1978; Shatz, 1981). Jacobson, Follette, and Revenstorf (1984) proposed a reliable change index (RCI), which was based on the SEM, as a means to evaluate psychotherapeutic change in individuals over time. The RCI was created to ensure that observed test score change is statistically reliable (one part of their criteria for clinically significant change). Reliable change refers to a difference in observed test scores that exceeds the amount of variation that could be reasonably attributed to measurement error. The RCI was originally defined as:

$$RCI = X_2 - X_1 / SEM \qquad \text{(Equation 3)}$$

As previously defined, $SEM = SD_X (1 - r_{xx})^{1/2}$ where SD_X is the pretest or normal control group standard deviation and r_{xx} is the (test-retest) reliability coefficient.

The use of the RCI assumes that the true score of the individual remains constant from time 1 to time 2. RCIs are based on a fixed-alpha strategy and therefore their

interpretation is similar to null hypothesis testing. After the alpha level is set, the critical z-score(s) are determined to mark the fixed boundaries of reliable change. For $\alpha = .05$ (two-tailed), the RCI must exceed 1.96 for the change to be deemed a statistically reliable improvement. A decrement in performance is identified as statistically reliable if the RCI is less than –1.96. RCI scores falling between these two critical cutoff points represent no reliable change; this amount of change is expected to occur by chance 95% of the time. A more lenient RCI of ±1.645 ($\alpha = 0.10$, two-tailed) is also commonly used in practice.

Speer (1992) attempted to improve Jacobson et al.'s (1984) RCI by correcting for the effects of regression to the mean. In accordance with the methods of Edwards et al. (1978) and Nunnally (1967), a regression adjustment was made to the numerator of the RCI by replacing the observed pretest score with an estimate of the individual's true initial score (which is always closer to the mean). The formula (labeled after Speer) is:

$$RCI_{SPEER} = (X_2 - (r_{xx} (X_1 - M) + M)) / SEM \qquad \text{(Equation 4)}$$

M = the mean test score in the general population. All other variables are as previously defined.

Interpretation of RCI_{SPEER} is similar to the original RCI. Speer (1992) recommended treating RCI scores above 2 as significantly improved and RCI scores below –2 as significantly deteriorated. One limitation of RCI_{SPEER} is that it does not account for practice effects. It has also been criticized for using an improper standard error term (i.e., the SEM) and for ignoring the unreliability inherent in the measurement of the posttest score (Hageman & Arrindell, 1993).

The RCI, as defined in most current research, no longer employs the SEM in the denominator (Jacobson & Revenstorf, 1988; Jacobson et al., 1999; Jacobson & Truax, 1991). The formula was amended following Christensen and Mendoza's (1986) suggestion that the standard error of difference (SED) between two observed test scores was the more appropriate error term. The SED refers to the distribution of difference scores that one would expect from the same person on the same test as a function of measurement error alone (i.e., when no real change has occurred). It has traditionally been operationalized as $(2\ SEM^2)^{1/2}$. Using this definition, the SED is always larger than the SEM (by a factor of 1.414) and it therefore results in a more stringent criterion for change. The new formula (labeled after Jacobson and Truax) is:

$$RCI_{JT} = (X_2 - X_1) / SED \qquad \text{(Equation 5)}$$

In light of recent confusion in the literature (see Abramson, 2000; Hinton-Bayre, 2000; Temkin, Heaton, Grant, & Dikmen, 2000), it should be noted that there are several methods for computing the SED. The most common method (Jacobson & Truax, 1991) mentioned above, simply involves multiplying the SEM by $\sqrt{2}$ (Equation 5a). This method provides an approximation of the SED since it assumes that the standard deviation of the test scores are equivalent at both time 1 and 2. This assumption may not be correct. The SED has alternatively been defined as $(SEM_1^2 + SEM_2^2)^{1/2}$ (Anastasi, 1988; Iverson, 1999), which takes into account the SEM at both baseline and re-test (Equation 5b). If longitudinal data are available, the SED may also be directly calculated (Equation 5c) as the standard deviation of the observed difference scores (Temkin, Heaton, Grant, & Dikmen, 1999; Temkin et al., 2000). To the best of the author's knowledge, the practical impact of using one method over another is unknown. Although

direct empirical measurement of the SED might ordinarily be preferred to a theoretical estimate, this issue has not been investigated in clinical research.

The RCI_{JT} is interpreted in a manner similar to the original RCI. Difference scores that exceed critical z-values multiplied by the SED are defined as statistically reliable change (e.g., $X_2 - X_1 > 1.96$ SED). The simplicity of the RCI_{JT} has made it popular in both the psychotherapy and neuropsychological literature (Hinton-Bayre et al., 1999; Iverson, 1999, 2000; Jacobson et al., 1999). Although the RCI_{JT} yields important categorical information (i.e., reliable improvement, no reliable change, or reliable decrement), it is not meant to explicitly measure the relative magnitude of individual change. Furthermore, it is not amenable for use in making comparisons among different measures since the index is expressed in the units of a specific measure. The RCI_{JT} does account for errors due to the unreliability of the measure, but it does not make specific adjustments for practice effects or regression to the mean (Hsu, 1989, 1995; Speer, 1992).

There have been several attempts to improve the RCI_{JT}. Chelune, Naugle, Luders, Sedlak, and Awad (1993) proposed a correction for the RCI_{JT} that accounts for practice effects. Their correction simply involves subtracting a constant value from the observed difference score. The constant is typically the mean amount of group improvement or decrement over a specified interval in a control sample. The formula (labeled after Chelune) is:

$$RCI_{CHEL} = ((X_2 - X_1) - (M_2 - M_1)) / SED \qquad \text{(Equation 6)}$$

M_1 and M_2 are the observed pretest and posttest means of a control group, respectively. As with the RCI_{JT}, the SED in Equation 6 can be estimated by multiplying the SEM by $\sqrt{2}$ (Equation 6a), defined as $(SEM_1^2 + SEM_2^2)^{1/2}$ (Equation 6b), or objectively determined

by using the standard deviation of the observed difference scores (Equation 6c). The interpretation of the RCI_{CHEL} is the same as other RCIs. It should be noted that the RCI_{CHEL} is a special case application of a formula specified by Payne and Jones (1957) for determining the reliability of a discrepancy between two scores.

The RCI_{CHEL} has been employed in several neuropsychological studies (e.g., Chelune et al., 1993; Hermann et al., 1996; Ivnik et al., 1999; Kneebone et al., 1998) and has been viewed as an appropriate means to measure individual change in cognitive abilities. The main limitation of this method is that practice effects associated with any specific measure are assumed to be uniform for all people. This assumption is likely invalid since practice effects, as previously mentioned, are also determined by the test-retest interval and the characteristics of the persons who comprise the reference sample (e.g., young versus old, cognitive impairment present versus absent).

Hsu (1989; 1999), like Speer (1992), proposed an alternate RCI formula to correct for the effects of regression to the mean. Hsu's modification involved replacing the observed difference score in the RCI_{JT} equation with a "residualized gain" score to take into account an individual's level of performance relative to the group mean. The residualized gain score in the numerator was viewed as an improved estimate for the true change score. The standard error term relevant to a residual change score is the standard error of prediction (SEP). Accordingly, the SED in the denominator of the RCI_{JT} was replaced with the SEP. The resulting formula (labeled after Hsu) is:

$$RCI_{HSU} = ((X_2 - M_2) - r_{xx}(X_1 - M_1)) / SEP \qquad \text{(Equation 7)}$$

Recall that $SEP = SD_X (1 - r_{xx}^2)^{1/2}$, M_1 = mean pretest score, and M_2 = mean posttest score. This method is similar to regression-based change score methods (to be discussed

later) and is a special case application (i.e., where $SD_X = SD_1 = SD_2$) of a formula originally described by Payne and Jones (1957) for testing a clinical prediction. The interpretation of the RCI_{HSU} is the same as the RCI_{JT}.

A major criticism leveled against the RCI_{HSU} method is that the relevant group mean to which test scores are supposed to regress toward may not be known or easily determined. Nunnally and Kotsh (1983) have addressed this issue and recommend using the general norms that exist for a specific measure when an individual's group membership is in question. Hageman and Arrindell (1993), in contrast, have suggested that reference need only be made to the observed pretest and posttest means. A second criticism against the RCI_{HSU} has been forwarded by Maassen (2000b) who distinguishes between classical null-hypothesis derived RCIs (i.e., RCI, RCI_{JT}, RCI_{CHEL}) and those RCIs, such as the RCI_{HSU}, that are interval estimation methods. Maassen (2000b) states that the latter methods identify an interval that most likely contains the true score difference and if this difference does not contain zero, then reliable change is inferred. However, interval estimation methods are not based on a uniform probability distribution (as with classical null-hypothesis derived RCIs) that would allow one to estimate the probability of making a Type I error. Moreover, he contents that interval estimation methods are biased estimates of the true score difference that are prone to increase misclassification errors for extreme test scores and low reliability baseline scores.

Hageman and Arrindell (1993; 1999a; 1999b) have proposed two different refinements of the RCI. The first, named RC_{ID} (for "improved difference" score), modifies the RCI_{JT} numerator substantially by accounting for regression to the mean due to measurement unreliability. The reliability term used to estimate measurement error is

the reliability of the difference score (r_{DD}). In the denominator, the SED term is retained but is calculated based on separate SEMs for the pretest and posttest. This differs from Jacobson and Truax's (1991) method in which a single SEM value is assumed for both the pre- and posttest score distributions (i.e., Equation 5a). For the calculation of the pretest and posttest SEMs, Hageman and Arrindell (1993) recommended the use of Guttman's (1945) reliability coefficients. These coefficients represent the lower bounds of the reliability of a measure calculated from a single sample. Accordingly, the formula for RC_{ID} is:

$$RC_{ID} = (r_{DD} (X_2 - X_1) + (1 - r_{DD}) (M_2 - M_1)) / (SEM_1^2 + SEM_2^2)^{1/2} \quad \text{(Equation 8)}$$

In this formula, $r_{DD} = SD_1^2 r_{xx(1)} + SD_2^2 r_{xx(2)} - 2 SD_1SD_2r_{xx} / SD_1^2 + SD_2^2 - 2SD_1SD_2r_{xx}$. $SEM_1 = SD_1 (1 - r_{xx(1)})^{1/2}$ and $SEM_2 = SD_2 (1 - r_{xx(2)})^{1/2}$. The values for $r_{xx(1)}$ and $r_{xx(2)}$ are the highest Guttman's reliability coefficients, r_{xx} is the test-retest correlation coefficient, and SD_1 and SD_2 are the standard deviations for the pretest and posttest scores, respectively. The RC_{ID} is interpreted the same as other RCIs. The arguments leveled against interval estimate methods (Maassen, 2000b) that were previously outlined apply to the RC_{ID}. In a separate article, Maassen (2000a) specifically addresses the RC_{ID} and argues that since its denominator does not contain the standard error term of the numerator, the index does not conform to a standardized normal distribution and is not amenable to judgments regarding the probability of making a Type I error.

The latest index from Hageman and Arrindell (1999a; 1999b), named RC_{INDIV}, is unique in that it does not employ a fixed-alpha strategy like other reliable change indices. It instead uses a phi-strategy introduced by Cronbach and Gleser (1959) in which the risk of being misclassified as "improved" or "deteriorated" is set to a maximum allowable

value (e.g., 5%). Cronbach and Gleser (1959) argued that any use of a cut-off point (e.g., z_α) results in an increased risk of misclassification for values nearer to that cut-off and proposed the phi-strategy to keep this risk of misclassification at a constant. There is an important distinction between the phi-strategy and the more popular alpha-strategy used in decision-making. The fixed-alpha strategy of the RCI_{JT} assumes that the true difference is zero (i.e., no real change from time 1 to time 2) and a sufficiently large RCI value allows one to reject this null hypothesis and infer that true change has occurred. The question addressed by the RCI_{JT} is therefore: "Given an individual for whom the true difference is zero, how likely is it that we will interpret a difference?" The RC_{INDIV} based on the phi-strategy answers a slightly different question: "Given an individual with an observed difference, how likely are we to be correct in classifying the difference?" (McGlinchey et al., 1999, p.212). An absolute value of $RC_{INDIV} > 1.65$ indicates statistically significant change at the individual-level with a maximum 5% chance of misclassifying the direction of change.

The RC_{INDIV} is similar to the RC_{ID} with one major exception. The denominator is changed from using the SEMs of the pretest and posttest scores to using a formula equivalent to the standard error of estimation (SEE) for estimating the true difference between pretest and posttest scores from the observed difference. The formula is:

$$RC_{INDIV} = (r_{DD} (X_2 - X_1) + (1 - r_{DD}) (M_2 - M_1)) / (r_{DD} \times 2\ SEM^2)^{1/2} \quad \text{(Equation 9)}$$

The calculation of r_{DD} is as before, but $r_{xx(1)}$ and $r_{xx(2)}$ are defined in terms of the SEM. Hageman and Arrindell recommended calculating only one SEM based upon the best available reliability coefficients (i.e., those obtained in a specific sample under optimal conditions). Appropriate reliability coefficients might be one of Guttman's reliability

coefficients, the alpha coefficient, or the test-retest coefficient (so long as no relevant change occurs between time 1 to time 2). These authors state that the application of RC_{INDIV} in actual practice should be limited to situations where $r_{DD} \geq 0.40$ as the index is very sensitive to underestimated values of r_{DD}. The RC_{INDIV} creators claim that it is more sensitive than other RCIs to declining scores but the use of the phi-strategy for decision-making is neither well-known nor widely applied. The utility of this approach needs to be adequately tested in clinical research.

To summarize, there are many indices of reliable change. The RCI_{JT} and the RCI_{CHEL} are the simplest and most common forms of the RCI. They have been employed in several neuropsychological studies including one investigation focused exclusively on an elderly sample. Ivnik et al. (1999) examined Mayo's Older American Normative Studies (MOANS) data from older adults who were assessed every one or two years on at least three occasions. These authors found that different cognitive areas (e.g., verbal comprehension and retention of information) have varying degrees of temporal stability in normal adults and therefore require different magnitudes of change to be considered reliable. Ivnik et al. (1999) did not specifically examine the diagnostic sensitivity of RCI-determined change. The other RCI methods are largely untested in the neuropsychological literature and comparisons among the methods are necessary to determine which, if any, are appropriate for clinical use with older adults.

Standardized regression-based change scores. Over the last decade, regression analyses have been used to generate norms for neuropsychological measures that correct for the influence of demographic factors such as age, gender, and education (Heaton, Chelune, Talley, Kay, & Curtiss, 1993; Tuokko & Woodward, 1996). Regression

analyses may also be employed to measure cognitive change at the individual-level as originally demonstrated by McSweeney, Naugle, Chelune, and Luders (1993). In this approach, persons from a control sample complete a neuropsychological battery on two separate occasions. The data from the control sample are used to generate a regression equation in which posttest scores are predicted from observed pretest scores (i.e., simple regression). Application of the regression equation allows one to generate an expected or predicted time 2 score for an individual based on his/her performance at time 1 (i.e., predicted X_2 = beta weight * X_1 + constant). Standardized regression-based (SRB) change scores (labeled after McSweeney) are calculated by the dividing the discrepancy between the expected score and the observed score at time 2 by the standard error of estimate (SEE) in the regression equation:

$$SRB_{MCS} = (X_2 - \text{predicted } X_2) / SEE \qquad \text{(Equation 10)}$$

The SEE in a multiple regression analysis is defined as $(SS_{residuals} / N - k - 1)^{1/2}$. Like the fixed interval that marks the boundary of real change in RCI equations, SRB_{MCS} change scores that exceed a specific value (e.g., $\pm z_{\alpha/2}$ SEE) raise the suspicion of real change.

SRB change scores may also be developed that account for more than simply the observed pretest score. Multiple regression, in contrast to simple regression, involves generating an equation that includes the pretest score in addition to any other relevant variables that may influence test performance. Age, education, gender are the variables that are most widely known to influence cognitive test performance. Other factors, however, such as overall cognitive status, emotional state, and medication use might also impact test performance. As above, application of the multiple regression equation allows one to generate an expected time 2 score for an individual (i.e., predicted X_2 =

(beta weight * X_1) + (beta weight * V_1) + … + (beta weight * V_n) + constant). Multiple regression-based change scores are calculated using the same formula as SRB_{MCS} and are interpreted in a similar fashion:

$$SRB_{MULT} = (X_2 - \text{predicted } X_2) / SEE \qquad \text{(Equation 11)}$$

Recently, Crawford and Howell (1998) introduced a new, and more technically accurate, SRB method for comparing predicted and obtained scores. This newer method addresses the error that arises from the use of sample coefficients to estimate population regression coefficients. In McSweeney et al.'s (1993) approach, the regression equation is specific to the sample and therefore represents an optimal fit of the sample data. It is assumed that the sample is representative of the population; that is, the derived equation may be used to accurately predict time 2 scores for individuals who were not in the original sample. This may not be true. The use of sample statistics, which fails to adjust the regression equation to reflect the estimation of population regression coefficients, may increase the likelihood of identifying discrepant scores as significantly changed. This error would be magnified for pretest scores that are further from the mean pretest score. The new method accounts for this potential error by multiplying a correction factor to the SEE for each individual case. The formula for the proposed correction (labeled after Crawford and Howell) is:

$$SEE_{CH} = SEE\ (1 + 1/N + ((X_1 - M_1) / SD_X^2\ (N-1)))^{1/2}$$

N = total number of persons in sample, SD_X is the standard deviation of the pretest scores, and all other variables are as previously defined.

When the new SEE is substituted into equation 9, it becomes:

$$SRB_{CH} = (X_2 - \text{predicted } X_2) / SEE_{CH} \qquad \text{(Equation 12)}$$

The authors recommended the use of the t-statistic, rather than the z-statistic, when working with samples rather than populations. A $t_{\alpha/2}$ ($df = N - 2$) is therefore used to replace the $z_{\alpha/2}$ value (e.g., 1.96 or 1.64) used in other methods to demarcate the bounds of reliable change.

Crawford and Howell (1998) employed hypothetical neuropsychological data to examine the impact of using the unadjusted and technically correct regression-based methods. Their examination suggested that the unadjusted method systematically yielded narrower confidence intervals than those obtained using the correct method. For sufficiently large sample sizes (i.e., $N > 100$) and pretest scores that were not extreme (i.e., > 2 SDs), the differences between the two approaches were modest. The authors recommend using the technically correct method with smaller samples. Crawford and Howell's (1998) correct method has been applied in clinical neuropsychological research (e.g., Graves, 2000) but has not yet been investigated with respect to change scores in older adults.

The strength of regression-based approaches in change measurement is that they control for practice effects, regression to the mean, and any other test-retest confound observed in the normal population for a particular measure (McSweeney et al., 1993). By factoring out the variance of the pretest score from the posttest score, this approach essentially serves to equate individuals who differ in their baseline performance. Another advantage is that regression-based change scores may be expressed as continuous variables in terms of a common metric (e.g., z-scores or T scores) thus facilitating comparison of scores among different measures. This differs from the limited

categorical information yielded by RCIs (i.e., reliable improvement, no reliable change, or reliable deterioration).

The SRB methods have many advantages over other change score methods but notable limitations also exist. The SRB methods described above are not appropriate when the assumptions of multiple regression are violated. The relation between the pretest and posttest scores should be linear and homoscedastic and the predictor(s) should be measured without error (Pedhazur, 1982). The assumption of classical test theory regarding the fallibility of measurement is inconsistent with the assumption underlying regression analysis. McSweeney et al. (1993) recommended that regression-based methods should not be used when the data for change are not normally distributed. As well, measures prone to floor or ceiling effects are not amenable for use with regression-based methods. Finally, one needs to consider the appropriateness of the regression equation for use with a specific individual. The accuracy of regression equations may be compromised when applied to individuals whose scores or characteristics are outside of the range of the reference sample from which the equation was derived. It is not clear how robust regression-based methods are to violations of these assumptions.

In the neuropsychological literature, regression-based methods have predominantly been used to study post-surgical cognitive change in individuals with epilepsy (McSweeney et al., 1993; Sawrie, Chelune, Naugle, & Luders, 1996). Recently, Sawrie and his colleagues (Sawrie, Marson, Boothe, & Harrell, 1999) extended the use of regression-based methodology to study individual cognitive decline in older adults. Their study involved examining the one year test-retest data of a small sample of 23 neurologically intact, community-dwelling, older adults (mean age = 66.5 years). The

neuropsychological battery included the Mattis Dementia Rating Scale (MDRS; Mattis, 1988), subtests from the Wechsler Adult Intelligence Scale - Revised (WAIS-R; Wechsler, 1981), subtests from the Wechsler Memory Scale - Revised (WMS-R; Wechsler, 1987), Trail Making Tests (TMT; Reitan & Wolfson, 1985), the Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 1983), and measures of letter and category fluency (Benton & Hamsher, 1978; Spreen & Strauss, 1998). Mean performance remained relatively stable for most measures over the study interval.

The neuropsychological and demographic data from the 23 study participants were used to generate regression equations that were then retrospectively applied to data from three persons diagnosed with differing forms of dementia. Change scores extending beyond the 90% confidence interval (i.e., ±1.64 SEE) were considered statistically rare and clinically relevant. In each case, the pattern of change detected using SRB methods was consistent with the dementia diagnosis. For example, the person with Alzheimer Disease evidenced significant deterioration on global cognitive, memory, and language measures. The individual with Pick's disease demonstrated pervasive deficits in memory, language, and executive functioning. Significant improvement in language functioning without evidence of decline in other domains was seen in a person diagnosed with vascular dementia. Though illustrative, the examination of three selected cases does not speak to the diagnostic sensitivity of the SRB methods to detecting dementia in a larger sample of persons with and without cognitive impairment. It should also be emphasized that the regression equations generated by Sawrie et al. (1999) were based on a small sample of older adults and as such, are not appropriate for generalization to the population of persons over age 65.

Comparing methods of change measurement

A variety of RCI and SRB methods have been proposed over the last decade to assist clinicians in determining the significance of changes observed in test performance over time. With each proposal, there has been considerable debate as to the "right" way to address errors and biases in measurement and the proper standard error term that should be used. It is surprising that few attempts have been made to directly compare these methods. This may, in part, reflect the fact that at least two of the methods have been introduced only recently (i.e., Crawford & Howell, 1998; Hageman & Arrindell, 1999b). Jacobson et al. (1999) acknowledged the current state of the literature and concluded "less mathematical wrangling and more empirical testing is needed" (p. 306) to determine the utility of different change scores.

Speer (1992; 1995) was the first to examine the relation among different RCIs. In his initial study (Speer, 1992), the RCI_{JT} and RCI_{SPEER} were compared using test-retest data from 92 participants on a scale of general well-being. He found that the methods were not dramatically different and produced slight, but insignificant, differences in rates of improvement and deterioration. In 1995, Speer compared RCI_{JT}, RCI_{HSU}, and RCI_{SPEER} with hierarchical linear modeling (HLM) using multi-wave data from 73 outpatients on the same scale. With the exception of the RCI_{HSU}, there was considerable agreement (ranging from 78% to 81%) among the various methods in terms of the proportion of cases classified as "improved" and "not improved." The HLM method was more likely than the other methods to classify a change in test scores as improved but failed to identify a single case as significantly deteriorated. The other methods were slightly more conservative and yielded similar classifications for reliable change. The

RCI_{HSU}, in contrast, had the lowest agreement with the other methods; it generated the lowest improvement rate and the highest deterioration rate. Speer (1995) favored the HLM method, but recommended use of the RCI_{JT} method in situations in which there are only two testing occasions.

Kneebone et al. (1998) examined two change score methods using neuropsychological test data. These researchers compared the RCI_{CHEL}, which corrects for practice effects, to the SD method in 50 patients following coronary artery bypass grafting. RCIs were calculated using a 90% confidence interval based on the initial and follow-up data of 24 control participants (7-day test-retest interval). Using the other method, post-operative change was considered to have occurred if a participant's change score was greater than or equal to 1 SD of the group mean baseline score on the measure. The neuropsychological battery included the California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987), Purdue Pegboard (Tiffin, 1968), word fluency measures (Benton & Hamsher, 1978), TMT Parts A and B (Reitan & Wolfson, 1985), Digit Symbol subtest from the WAIS-R (Wechsler, 1981), and the BNT (Kaplan et al., 1983). Test-retest reliability coefficients over the one-week interval ranged from 0.67 to 0.94 and significant practice effects were found on the TMT, Digit Symbol subtest, and BNT. The RCI_{CHEL} method classified more patients as showing significantly more post-operative decline than the SD method on 5 of the 11 neuropsychological measures (including Purdue Pegboard, TMT Part B, BNT, and the Digit Symbol subtest). The SD method classified more individuals as deteriorated than the RCI_{CHEL} on the three CVLT indices that were examined, although the differences between the change score methods were not statistically significant. The investigators interpreted these findings as evidence

of the superiority of the RCI_{CHEL} over the SD method as it accounted for both practice effects and measurement unreliability.

Bruggemans et al. (1997) also examined neuropsychological test data from persons who had undergone cardiac surgery. These investigators compared the SD, RCI_{JT}, RCI_{CHEL}, and SRB_{MCS} methods using data from a sample of 63 patients seen over four occasions. In addition, they included a complex method for measuring change that involved controlling for error and practice effects by matching each patient with a group of control participants on the basis of pretest scores. With the exception of the SD method, critical values for determining reliable deterioration were based on $z > 1.645$ ($\alpha = 0.05$, one-tailed) for all methods. The battery of measures included the Rey Auditory Verbal Learning Test (RAVLT; Lezak, 1995; Rey, 1964), subtests from the WMS-R (Wechsler, 1987), word fluency (Benton & Hamsher, 1978), TMT (Reitan & Wolfson, 1985), the Stroop Interference test (Stroop, 1935), and the Symbol Digit Modalities Test (Smith, 1982). Measures of verbal fluency, attention and psychomotor speed were highly reliable ($r = 0.76$ to 0.92) and were associated with significant practice effects, whereas the learning and memory measures had lower reliability coefficients ($r = 0.45$ to 0.79) and no practice effects. In the learning and memory measures, the use of the SD method (which does not correct for measurement error or the effects of practice) resulted in an overestimation of deterioration rates relative to the other methods, which tended to be more conservative. There were few differences among the two RCIs and the SRB change scores under these conditions. For highly reliable measures, the failure to correct for practice effects (using either the RCI_{CHEL} or SRB_{MCS}) resulted in an underestimation of deterioration rates using the SD method and the RCI_{JT}. The authors concluded that the

various indices showed marked differences in deterioration rates consistent with their mathematical differences. Low reliability measures tended to show greater discordance between the SD and the other change methods and practice effects, when present, decreased the accuracy of methods that did not account for this bias.

Finally, Temkin, Heaton, Grant, and Dikmen (1999) compared RCI_{JT}, RCI_{CHEL}, SRB_{MCS} (simple linear regression), and SRB_{MULT} (stepwise multiple regression) change scores using two-wave neuropsychological data from 384 neurologically stable adults. The sample included 37 adults over the age of 65 years. Test-retest intervals varied substantially from 2.3 to 15.8 months (mean = 9.1 months). A total of 7 neuropsychological measures were examined including the Verbal IQ (VIQ) and Performance IQ (PIQ) from the original WAIS (Wechsler, 1955) and the Category Test (number of errors), Tactual Performance Test (total time), TMT Part B, the Halstead Index, and the Average Impairment Rating from the Halstead-Reitan Neuropsychological Test Battery (HRB; Reitan & Wolfson, 1993). Temkin et al. (1999) evaluated the four change score methods on the basis of 1) the width of the prediction interval yielded by each method, and 2) the accuracy with which each model fit an expected normal distribution of scores (in which 5% of cases were expected to show a significant improvement and 5% a significant deterioration). The Category Test and PIQ were associated with relatively large practice effects though these were not explicitly tested for statistical significance. Test-retest correlations were not provided but baseline performance was found to be the strongest predictor of follow-up performance across all measures. In comparing the various methods, the authors found that the RCI_{JT} was the least accurate since it consistently yielded the widest prediction intervals and classified

more cases than expected as improved and fewer cases as deteriorated. The RCI_{CHEL} yielded equally wide prediction intervals as the RCI_{JT} but improved the classification accuracy rate. The best method by their evaluation criteria was the SRB_{MULT} method, which incorporated baseline test performance and additional variables such as age, education, race, length of retest interval, and global impairment ratings. But overall, the SRB_{MULT} did not substantially improve classification accuracy over that of either the SRB_{MCS} or RCI_{CHEL}.

The comparison studies reviewed above differ in terms of the measures that were used, the sample studied, the test-retest interval, and the specific change methods that were examined. It is therefore difficult to draw firm conclusions, but it may be worthwhile to comment on the emerging patterns and trends. The findings across the studies suggest that the SD method is inferior to the RCI and SRB methods, particularly on lower reliability measures. This is consistent with what one might expect based on measurement theory since the SD method employs an arbitrary criterion for change and fails to account for a single error or bias. The RCI_{JT} represents an improvement over the SD method by accounting for measurement error but lacks the improved accuracy that comes with adjustments for practice effects. The RCI_{CHEL} has resulted in classification accuracy rates that are comparable to the more complex regression-based models (Heaton, 1999) though the SRB_{MULT} appears to yield the best results overall.

A major limitation of the studies that have compared methods for measuring change involves their evaluative standards. The common standard among these studies has been the proportion of cases classified as improved or deteriorated by each method. However, simply looking at the classification differences between two methods (e.g.,

Kneebone et al., 1998) does not adequately address whether either method actually captured "real" change (unless it is assumed that all cardiac surgery patients show substantial declines across a broad range of cognitive abilities). Comparing the observed improvement and deterioration rates to the pattern that is expected on the basis of chance alone (e.g., 5% to show improvement and 5% to show deterioration) also provides only limited evidence in support of a method's usefulness in change measurement. Consider Temkin et al.'s (1999) study. How much can be learned about the utility of different methods for detecting cognitive change when the sample is comprised of "neurologically stable" individuals who presumably did not evidence "real" change? A more rigorous test of a method's utility might involve examining a group of individuals for whom clinically significant change has been well documented. Of course, clinically significant change could be operationalized in a variety of ways; but the point is that a reasonable index of clinically meaningful change would certainly be a more appropriate benchmark for evaluating change methods than what has been employed to date.

The real world significance of statistically reliable change is an important issue that has not been thoroughly investigated (Beutler & Moleiro, 2001; Kazdin, 1999). The RCI_{JT}, for example, assesses whether an observed change exceeds that which might be reasonably attributed to measurement error but it does not specifically address how rare or abnormal a difference score is in the population. Research is needed to determine exactly what reliable change (as defined using different methods) on one or more neuropsychological tests actually means. Is reliable change related to a diagnostic change? Or is reliable change common among certain populations? Adopting clinical significance as a comparative standard may help answer these questions. Such a standard

not only provides a means by which to compare methods for measuring change, it may also serve to establish the validity of neuropsychological instruments for the purpose of measuring change. A frequent finding of reliable change on certain tests in the absence of clinically meaningful change or a failure to observe reliable change when clinically significant change was deemed to have occurred may suggest that the instrument is inappropriate for change measurement.

Ivnik et al. (2000) provide some data regarding the relation between change scores and a clinically significant standard using data from the Mayo's Older Americans Normative Studies (MOANS). These investigators compared the diagnostic accuracy of RCI_{CHEL} using longitudinal neuropsychological data to various cross-sectional methods for assessing cognitive deficits in a sample of older adults. Five Mayo Cognitive Factor Scores (MCFS) based on WAIS-R (Wechsler, 1981), WMS-R (Wechsler, 1987), and RAVLT (Rey, 1964) data served as the dependent neuropsychological variables. The MCFS include an index of verbal comprehension, perceptual organization, attention/concentration, learning, and retention. Participants were seen on at least two occasions and at each visit were classified as either normal or cognitively impaired by a multidisciplinary group of health care professionals. The latest consensus diagnostic opinion regarding each individual served as the "gold standard" for the evaluation of diagnostic accuracy. The cognitively impaired group was heterogeneous and included persons with dementias (e.g., Alzheimer Disease, vascular dementia, mixed dementia) and mild cognitive impairment.

The cross-sectional methods that were examined in this study included an absolute approach (using a single cut-off score to define cognitive deficit), a difference

approach (examining differences among various cognitive indices at baseline such as the difference between memory performance and reading ability), and a profile variability approach (based on the amount of variability in test performance at baseline). In the change score approach, change on each of the 5 MCFS indices was examined from Time 1 – Time 2, Time 1 – Time 3, and the total number of measures in the test battery for which reliable change was evident. A total of 1079 participants (672 controls and 407 with cognitive impairment) provided data for the cross-sectional methods and 331 persons (194 controls and 137 with cognitive impairment) provided test-retest data for examining the utility of the change score approach.

At the individual-level, the diagnostic accuracy for classifying persons as cognitively impaired or normal using the MCFS was highest for the absolute approach (range = 58% to 84% accuracy) suggesting that clinicians rely more heavily on the absolute test values than on comparisons within or between test scores. It was expected that cognitively impaired persons would show more change on test scores than cognitively normal older adults, but contrary to their hypothesis, the observed change scores rarely exceeded chance levels for distinguishing between the groups in this study (range = below 50% to 68% accuracy). Similar results were obtained when analyses were performed on a more restrictive sample using data from normal participants and persons with Alzheimer Disease (rather than the more heterogeneous "cognitively impaired" group). The authors interpreted these findings as evidence that the change score approach was not diagnostically useful. They speculated that normal age-related variability in test performance of older adults may be greater than previously appreciated

making it difficult to identify clinically significant decline in persons with cognitive impairment.

Some points regarding Ivnik et al.'s (2000) study should be highlighted. First, the conclusion regarding the utility of change scores with older adults might be inappropriate as only one change score method was examined. Second, it can be argued that the design of the study did not permit an adequate test of the utility of the change score approach. Ivnik et al. (2000) identified two distinct groups for the change score analyses based on the most recent consensus diagnosis - one with cognitive impairment and one without. The difficulty with this design is that persons identified with cognitive impairment may not have been "normal" at baseline and may not have evidenced a real change. Many of the participants may have had stable cognitive impairments (e.g., mild cognitive impairment at time 1 and time 2) in which case little change in test performance would be expected. Alternatively, several of the participants could have been diagnosed with dementia at time 1. If this were true, it is uncertain what diagnostic value a clinician would derive from examining a change score in someone whose dementia status was already known at baseline. It has been argued here that a more stringent test of the diagnostic utility of change scores would involve examining persons whose cognitive status was known to be normal at baseline and then actually exhibited a diagnostic change over the study interval. A third consideration pertaining to Ivnik et al.'s (2000) study relates to the influence of floor effects. The investigators failed to rule out the possibility that floor effects associated with the MCFS may have prevented change scores from adequately discriminating between persons deemed to be cognitively intact and impaired. The cognitively intact group, not surprisingly, started with higher average

scores on the five MCFS (range = 100.2 to 105.9) than the cognitively impaired group (range = 79.2 to 97.2). The lower scores in the cognitively impaired group may have allowed little room for detecting additional loss. Closer inspection of the time 1 - time 2 data support this possibility. Mean change in the normal group ranged from 1.3 to 6.9 points on the five MCFS whereas mean change for the cognitively impaired group was substantially less ranging from -0.6 to -1.3 points. The above cited points limit the conclusions that may be drawn from this study regarding the utility of change score methods and support the need for future research to address this issue.

Recently, Heaton et al. (2001) extended their previous findings (i.e., Temkin et al., 1999) to explore the sensitivity of change score methods to "real" change using clinical samples. In contrast to Ivnik et al. (2000), the findings from this study suggest that change score methods are useful in the determination of clinically significant change. These authors examined the RCI_{CHEL}, SRB_{MBS}, and SRB_{MULT} in four new participant groups including a non-clinical cross-validation group (n = 124), a stable group of persons with chronic schizophrenia (n = 69), a group of persons recovering from moderate to severe traumatic brain injuries (TBI; n = 23), and a group of individuals with new brain insults (n = 10). The latter two groups comprised persons with clinically significant improvement and deterioration, respectively. The small group of persons with new brain insults included 5 persons with cerebrovascular accidents, 1 TBI patient who required surgical evacuation of a subdural hematoma, 2 persons with complications related to HIV infection, and 2 persons with Alzheimer Disease. The mean test-retest interval was 11 months for the "improved" group and 35 months for the "deteriorated" group.

Using similar procedures as those described in Temkin et al. (1999), the authors found that the change score methods were not accurate in classifying "no change" in the stable group of persons with schizophrenia suggesting that these methods, in conjunction with the use of the WAIS and HRB, lack specificity. More promising results were obtained in the groups with clinically significant change. The RCI_{CHEL} correctly classified 74% of the recovering TBI group as reliably improved and 80% of the new insult group as showing reliable deterioration. The SRB_{MCS} resulted in comparable accuracy to the RCI_{CHEL} but was slightly better, with 90% accuracy, in classifying decline in the new insult group. The full regression model, SRB_{MULT}, was less accurate than the other two methods having correctly classified 65% of the improved group and 70% of the deteriorated group. The best accuracy rates were obtained when all 7 neuropsychological measures in the study were combined rather than relying on any one particular measure.

The authors concluded that norms for change from non-clinical samples may not generalize well to clinical groups, such as persons with schizophrenia. However, their norms for change resulted in accurate classification using all three change score methods in samples who were identified as having demonstrated clinically significant change. In contrast to Ivnik et al. (2000) who found little support for using change score methods, Heaton et al. (2001) concluded that the relatively simple RCI_{CHEL} was not only accurate, but comparable to the more sophisticated SRB methods in both normal and clinical samples. While these findings suggest that change score methods may hold promise in detecting and classifying diagnostic change in samples of persons with declining cognitive abilities, Heaton et al. (2001) did not specifically examine change in older

adults or in persons who developed dementia. Moreover, their examination of change was restricted to the WAIS and the HRB, which do not adequately assess important cognitive domains such as memory functioning.

Rationale and description of the study

The examination of statistically reliable change in neuropsychological performance has predominantly focused on persons with epilepsy, persons with acquired brain injury, or cardiac patients. There have been relatively few efforts to investigate the reliable change that occurs as part of the normal and pathological changes associated with aging. This study focused on adults over the age of 65 and was designed to extend the findings of previous researchers (i.e., Sawrie et al., 1999; Temkin et al., 1999; Heaton et al., 2001; Ivnik et al., 1999, 2000) by examining different methods for measuring neuropsychological change in this population. The present study was also intended to add to the body of research regarding the suitability of change scores that, to date, has produced mixed results. Although change scores are being used more often in the neuropsychological literature, researchers have not yet adequately assessed the variety of change scores that currently exist.

The primary aim of the study was to determine which neuropsychological measures (if any) and statistical methods (if any) are appropriate for measuring change in older adults. The main research questions were: 1) How much change in neuropsychological test performance is normal, as defined by various change score indices, among older adults without cognitive impairment over (a) a relatively brief test-retest interval and (b) a relatively long test-retest interval?; and 2) Which statistical method(s), if any, are accurate in detecting clinically significant cognitive change in older

adults at the individual-level? The identification of variables and statistical methods that might improve the accuracy and suitability of the change methods in clinical use underscore the value of this study, particularly if change score methods can help clinical neuropsychologists to detect dementia in its earliest stages.

The first research question refers to the extent to which change occurs in the neuropsychological performance of cognitively intact older adults. Test-retest data were employed to examine change at both the individual and group levels of analyses. Group-level change was examined by calculating the mean difference score from pretest to posttest (i.e., $X_2 - X_1$). This provided an indication of the practice/aging effect associated with each measure. Individual change was investigated with methods similar to those Temkin et al. (1999) used in their examination of a cognitively stable group. Different techniques for measuring reliable change were compared in terms of the width of the prediction intervals and the proportion of cases classified as reliably improved, not reliably changed, or reliably deteriorated. The SD method (Equation 2), 7 RCIs (Equations 3 through 9), and the 3 regression-based methods (Equations 10 through 12) for calculating statistically reliable change were examined. While a small number of studies have examined a few select methods, an empirical comparison of all methods including the latest to emerge in the literature has not previously been attempted.

The current study comprised several phases. Phase A involved older adults with no cognitive impairment and examined variability in neuropsychological test performance over a relatively brief test-retest interval of approximately 3 months. The purpose of this phase was to quantify measurement error and obtain information about changes in test performance that occur during a span of time in which no true change in

cognitive ability was expected to occur. The results of Phase A provided information regarding the short-term reliability of various cognitive measures used with older adults as well as an indication of the group practice effects over a relatively short interval. Phase A may be viewed as an opportunity to replicate and extend the findings of Temkin et al. (1999) regarding the value of different change methods at the individual-level.

As a group, it was hypothesized that cognitively intact older adults would show a practice effect on all neuropsychological measures over this relatively brief interval. The effects of practice were expected to be greatest on timed tasks with a speeded component, as shown by previous research. To a lesser extent, practice effects were also expected on non-speeded tasks due to previous exposure to the test questions and familiarity with the testing situation. The effects of aging over an interval of a few months were not expected to be substantial and therefore were not expected to reduce practice effects in this group.

At the individual-level, it was expected that reliable changes in test performance would only occur at a rate consistent with chance misclassification reflected by the selected alpha level (e.g., 5% improvement and 5% deterioration with $\alpha = .10$) if the residuals (i.e., observed test score differences) for the measure were normally distributed. Methods for determining reliable change were expected to vary in their classification accuracy rate since each method accounts for error and bias in different ways. The statistical methods that account for the most error variance or bias were hypothesized to be superior (indicated here as ">") to those methods that account for fewer sources of error and/or bias. The specific hypotheses based on theory and previous research findings were that SRB_{CH} and $SRB_{MULT} > SRB_{MCS}$, RC_{INDIV}, RC_{ID}, and $RCI_{CHEL} > RCI_{HSU} > RCI_{JT} > RCI_{SPEER} >$ RCI > SD method. As previously mentioned, the SED

used in the calculation of RCI_{CHEL} and RCI_{JT} has been defined in at least 3 different ways but the impact of varying this definition has not been thoroughly examined. It was expected that the use of the observed SED from longitudinal data (i.e., Equations 5c and 6c) would yield better classification accuracy than methods in which the SED was estimated (i.e., Equations 5a, 5b, 6a, 6b).

The various methods for determining reliable change were anticipated to differ in terms of the width of the generated prediction intervals (i.e., the bounds of normal change). The regression-based methods were hypothesized to yield narrower prediction intervals than the RCI methods, which in turn would be superior to the SD method. Notably, the SRB_{CH} (Equation 12) could not be compared to the other statistical methods in this manner. Rather than generating a fixed width interval, the SRB_{CH} yields prediction intervals whose width is a function of the individual's baseline performance. As a result, the prediction interval is different for each participant and not amenable to straightforward comparison.

As a pattern, it was expected that no individual in Phase A would demonstrate reliable changes that were consistent with the pattern of cognitive impairments that define dementia. That is, no individual in Phase A, which included only persons who were cognitively healthy, was expected to show statistically reliable deterioration in memory and one or more other cognitive domains over a relatively brief interval.

In Phase B, the normal variability in older adults' neuropsychological test performance over a longer test-retest interval was investigated. The purpose of this phase was to examine the combined effects of measurement error, practice effects, and their interaction with normal aging in a sample of persons without cognitive impairment over a

span of several years. In this phase, aging effects might lead to true score changes in any given individual. The results of Phase B provided information regarding the long-term reliability of specific cognitive measures that could be directly contrasted with the findings from Phase A to determine the influence of varying test-retest interval length. It also allowed an examination of practice effects over a longer and more clinically relevant period of time. Phase B involved the same investigation of the various statistical techniques for determining reliable change described above in Phase A.

At the group-level, it was hypothesized that cognitively intact older adults, on average, would not demonstrate significant improvement on neuropsychological test performance over an interval of several years. Rather, mean scores were expected to remain stable or evidence a slight decline. Stability was expected if the combined effects of practice and aging cancelled each other out whereas decline would occur if aging effects, on the whole, were stronger than those associated with practice. For comparisons at the individual-level, the pattern predicted in Phase B was the same as that delineated in Phase A. Classification accuracy relative to the normal distribution of change scores was hypothesized to follow in descending order from SRB_{CH} and $SRB_{MULT} > SRB_{MCS}$, RC_{INDIV}, RC_{ID}, and $RCI_{CHEL} > RCI_{HSU} > RCI_{JT} > RCI_{SPEER} > RCI >$ SD method. Greater accuracy rates were expected for the RCI_{JT} and RCI_{CHEL} using formula c (the observed SED) relative to other methods for calculating the SED. As in Phase A, the widths of the prediction intervals were hypothesized to be narrower for the two regression-based methods relative to the RCI methods, which were expected to be superior to the SD method. Finally, no statistical method was expected to show a pattern of deterioration in

memory and one or more other areas of cognition as measured using neuropsychological tests.

The second research question pertained to the best method for detecting individual change in older adults that is abnormal or associated with a pathological process (i.e., change that is not due to aging alone). Phases C through G addressed the second research question in different ways and each phase involved an examination of the various methods for determining reliable change. A major advantage of this study over existing research (i.e., Ivnik et al., 2000) was the inclusion of cognitively healthy individuals who actually exhibited significant cognitive change. That is, persons who initially had no cognitive impairment at baseline and who had been diagnosed with dementia at follow-up approximately 5 years later. Individuals who exhibited this type of cognitive decline were the exclusive focus of Phase C.

Phase C may be viewed as the compliment to Phase B where the lack of reliable change was the expected finding in persons who did not evidence a change in diagnosis from initial testing to follow-up. The purpose of Phase C was to examine reliable change, as defined by the different methods, in the group of persons with known cognitive decline. Here, change in consensus diagnostic status was treated as evidence of a clinically significant change. For each statistical method, it is hypothesized that all persons who had no cognitive impairment at baseline and subsequently developed dementia would demonstrate reliable change on measures tapping into memory functioning and at least one other cognitive domain (e.g., language, abstract reasoning, constructional abilities). The results of Phase C analyses provided information about the validity of the statistical methods and the measures employed to determine change in

older adults in relation to a "gold standard." For example, the failure of a statistical method to detect a pattern of reliable change in two cognitive domains in a person diagnosed with dementia might suggest that either the method and/or the measures used to operationalize change in the specific cognitive domain are inadequate.

Phase D represented an attempt to overcome the shortcomings of previous research (e.g., Ivnik et al., 2000) by examining the relation between statistically reliable change and diagnostic change from a different perspective. The diagnostic accuracy of the various methods for determining reliable change was assessed by exclusively focusing on persons who were classified as cognitively healthy at baseline and either remained cognitively intact at follow-up or received a diagnosis of dementia. It has been argued here that this design provides a better benchmark of clinically significant change than the methods that have been employed in other studies to date. For example, other studies, which have included persons with "mild cognitive impairment" (i.e., an intermediate state between healthy aging and dementia), may not have captured the full range of change involved in the progression to dementia as individuals in the group may have already exhibited a substantial amount of cognitive decline. At the group-level, it was hypothesized that the group of persons who progressed to dementia would evidence greater mean change in all neuropsychological test scores relative to persons who remained cognitively intact. Analyses of change at the individual-level were expected to show the superiority of certain statistical change techniques over others in terms of classification accuracy. For each measure, SRB_{CH} and $SRB_{MULT} > SRB_{MCS}$, RC_{INDIV}, RC_{ID}, and $RCI_{CHEL} > RCI_{HSU} > RCI_{JT} > RCI_{SPEER} > RCI >$ SD method.

Reliable change on a single neuropsychological measure may be associated with a dementia diagnosis, but a pattern of change across several measures may be more important in clinical settings (Heaton et al., 2001). Therefore, Phase D also involved an examination of the number of measures on which each individual with no cognitive impairment at baseline demonstrated reliable change using each of the statistical methods. Persons who remained cognitively healthy or progressed to dementia formed two groups in order to examine the degree of association between diagnostic change and the sum of reliable change scores. It was hypothesized that persons with a large number of reliably changed scores would be more likely to develop dementia than persons who remained cognitively intact at follow-up.

In Phases E through G, the meaning of statistically reliable change was explored from perspectives other than diagnostic change. This was accomplished by comparing reliable deterioration as determined by the various statistical change methods to dichotomous ratings of cognitive change made by the individual (Phase E), an informant providing collateral information about the individual (Phase F), and health care professionals who examined the individual in a clinical evaluation (Phase G). To the best of the author's knowledge, an attempt to examine the meaning of statistically reliable change has not previously been attempted in the neuropsychological literature. The exploratory nature of this portion of the study precluded the formation of well-formed hypotheses. However, it was expected that statistically reliable deterioration in test scores would be strongly associated with ratings of loss provided by informants and mental health professionals across a range of measures and change methods. Ratings of change provided by the individual were expected to have a little or no relation to reliable

change on neuropsychological test scores. This latter hypothesis is consistent with research findings demonstrating a weak link between objective cognitive deficits and self-reported memory difficulties (Bolla, Lindgren, Bonaccorsy, & Bleeker, 1991; Feher, Larrabee, Sudilovsky, & Crook, 1994; McGlone et al., 1990; O'Connor, Pollitt, Roth, Brook, & Reiss, 1990; Sunderland, Watts, Baddeley, & Harris, 1986).

Hypotheses

The hypotheses that were tested in this study are explicitly listed below.

Group-level analyses of change.

1. In Phase A, which involved a relatively brief test-retest interval with cognitively intact participants, mean time 2 performance was expected to be greater than mean time 1 performance for all neuropsychological measures. All measures were expected to show some practice effects with more pronounced effects hypothesized to occur on speeded neuropsychological measures.
2. In Phase B, which involved a relatively longer test-retest interval with cognitively intact participants, mean time 2 performance was not expect to be greater than mean time 1 performance. All measures were expected to remain stable or show a group mean decline reflecting the mitigation of practice effects by normal aging.
3. In Phase D, which involved the examination of persons with no cognitive impairment at baseline, the group of individuals who progressed to dementia was expected to evidence greater mean decline in test performance across all neuropsychological measures relative to the group who remained cognitively intact.

Individual-level analyses of change.

4. In Phase A, it was hypothesized that SRB_{CH} and SRB_{MULT} would provide the closest fit to having 5% of the sample classed as "reliably deteriorated" and 5% of the sample classed as "reliably improved"; the SRB_{MCS}, RC_{INDIV}, RC_{ID}, and RCI_{CHEL} would be the next best group of methods in terms of classification accuracy followed by RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and the SD method. The $RCI_{JT}c$ was expected to be more accurate than $RCI_{JT}a$ and b methods and the $RCI_{CHEL}c$ was expected to be more accurate than the $RCI_{CHEL}a$ and b methods.
5. In Phase A, the regression-based methods (i.e., SRB_{MULT} and SRB_{MCS}) were expected to yield narrower prediction intervals than the 7 RCI methods, which were expected be superior to the SD method.
6. In Phase A, which involved cognitively intact participants, no statistical method was expected to show a pattern of reliable deterioration on one or more memory measures in addition to reliable deterioration on one or more neuropsychological measures assessing a different cognitive domain at a rate greater than chance.
7. In Phase B, it was hypothesized that SRB_{CH} and SRB_{MULT} would provide the closest fit to having 5% of the sample classed as "reliably deteriorated" and 5% of the sample classed as "reliably improved"; the SRB_{MCS}, RC_{INDIV}, RC_{ID}, and RCI_{CHEL} were expected be the next best group of methods in terms of classification accuracy followed by RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and the SD method. The $RCI_{JT}c$ was expected to be more accurate than $RCI_{JT}a$ and b methods and the $RCI_{CHEL}c$ was expected to be more accurate than the $RCI_{CHEL}a$ and b methods.

8. In Phase B, it was hypothesized that the regression-based methods (i.e., SRB_{MULT} and SRB_{MCS}) would yield narrower prediction intervals than the 7 RCI methods, which would be superior to the SD method.
9. In Phase B, which involved cognitively intact participants, no statistical method was expected to show a pattern of reliable deterioration on one or more memory measures in addition to reliable deterioration on one or more neuropsychological measures assessing a different cognitive domain at a rate greater than chance.
10. In Phase C, it was hypothesized that all persons who progressed to dementia would evidence reliable deterioration, using each of the statistical methods, on at least one memory measure and one other measure from a different cognitive domain.
11. In Phase D, it was hypothesized that diagnostic classification accuracy would be greatest for SRB_{CH} and SRB_{MULT} followed by SRB_{MCS}, RC_{INDIV}, RC_{ID}, and RCI_{CHEL} for each measure. Diagnostic accuracy was expected to progressively decrease for the RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and the SD methods.
12. In Phase D, it was hypothesized that the diagnostic accuracy of the sum of reliable changes would be greatest for SRB_{CH} and SRB_{MULT} methods followed by the SRB_{MCS}, RC_{INDIV}, RC_{ID}, and RCI_{CHEL} methods. Diagnostic accuracy was expected to progressively decrease for the RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and SD methods.
13. In Phase E, it was hypothesized that participants' subjective reports of loss in cognition would not be strongly associated with deterioration as indicated using the 3 regression-based methods, 7 RCI methods, and the SD method.

14. In Phase F, it was hypothesized that informants' reports of loss in cognition in the participant would be strongly associated with deterioration as indicated using the 3 regression-based methods, 7 RCI methods, and the SD method.
15. In Phase G, it was hypothesized that clinicians' rating of loss in cognition in the participant would be strongly associated with deterioration as indicated using the 3 regression-based methods, 7 RCI methods, and the SD method.

Method

Participants

All participants in this study were involved in at least one wave of the population-based Canadian Study of Health and Aging (CSHA). The CSHA is a large, multi-centre, multi-disciplinary, epidemiological study of health issues including dementia in people over age 65 in Canada (for more details, see Canadian Study of Health and Aging Working Group, 1994, 2000). The first wave of the CSHA (CSHA-1) began in 1991, the second wave (CSHA-2) in 1996, and the third and final wave (CSHA-3) in 2001.

In CSHA-1, a total of 10,263 persons from the community and institutions were interviewed regarding health-related issues and were screened for cognitive impairment using the Modified Mini-State Examination (3MS; Teng & Chui, 1987). All participants in institutions, all community participants scoring below 78/100 on the 3MS, and a subset of community participants scoring 78 or greater on the 3MS were seen for thorough clinical evaluations (n = 2914). The clinical component consisted of a nurse's evaluation, a physical examination, laboratory blood work, and a neuropsychological assessment.

The nurse's evaluation included a re-administration of the 3MS; a record of vision, hearing, vital signs, height, weight, and medication use; and the collection of informant-reported information (usually from a family member) about the participant's medical history, cognitive functioning, and functional abilities using Section H of the Cambridge Examination for Mental Disorders (CAMDEX; Roth, Huppert, Tym, & Mountjoy, 1988). The physical examination was conducted by a physician and involved an evaluation of the participant's head, neck, limbs, chest, and cardiovascular system in addition to primitive and central reflexes, peripheral neuromuscular responses, and coordination. Laboratory blood work was done for those participants with suspected dementia or delirium. For the neuropsychological evaluation, a trained psychometrician administered a standardized test battery to those participants who scored 50 or more on the 3MS from the nurse's evaluation. The neuropsychological test results were interpreted by a neuropsychologist. On the basis of all available information, CSHA clinicians (including physicians, psychologists, and nurses) met during a consensus conference and identified each CSHA participant as having one of the following: no cognitive impairment (NCI), cognitive impairment but no dementia (CIND), or dementia using DSM-III-R criteria (American Psychiatric Association, 1987).

CSHA-1 participants who completed these clinical evaluations were revisited approximately 5 years later during CSHA-2 using similar medical, psychological, and diagnostic procedures. Attrition due to death (n = 1534) or other loss of contact (n = 231) reduced the number seen for clinical evaluation at CSHA-2 to 1149 participants. In CSHA-3, all participants without dementia at CSHA-2 (i.e., those identified with NCI or CIND) were re-visited. An invitation to complete the CSHA-3 neuropsychological

evaluation was extended to all who had previously completed a clinical assessment at either CSHA-1 and/or CSHA-2 as well as to those individuals who had scored below 90/100 on the 3MS during CSHA-3 screening.

Retrospective data collected as part of CSHA-1 and re-test data collected during CSHA-2 were used in Phases B through G. The inclusion criteria for these retrospective analyses were: 1) the completion of the CSHA-1 neuropsychological component (English version), and 2) classification as having NCI at CSHA-1. Persons with CIND status (who by definition did not meet criteria for dementia) were not selected so as to minimize the risk of including persons who had already experienced a significant degree of cognitive impairment. Application of these criteria to the CSHA dataset resulted in the identification of 576 individuals. Data from one participant who had been diagnosed with mental retardation were removed leaving 575 eligible NCI participants.

During retest at CSHA-2, 208 out of 575 participants were again identified as having NCI. Eighty-eight participants were identified as CIND, and 49 as having a dementia. A total of 49 refused the study, 21 were lost to follow-up (inaccessible or not contacted), and 160 had died prior to CSHA-2. Of the 208 participants identified with NCI at both CSHA-1 and CSHA-2, 166 completed the neuropsychological battery at CSHA-2 and 42 did not. Twenty of the 49 participants with dementia at CSHA-2 completed the neuropsychological battery at follow-up.

Phase A, the prospective portion of this study, involved CSHA participants from Vancouver Island who were identified as NCI at both CSHA-2 and CSHA-3 (by high screening scores on the 3MS or consensus opinion) and who had completed the neuropsychological component at CSHA-3. CSHA-3 participants seen in Phase A had

not previously completed the neuropsychological assessment in CSHA-1 or CSHA-2. A letter and brochure (see Appendix A) were sent to CSHA-3 participants who met these criteria inviting them to participate in a follow-up using similar procedures to those used in CSHA-3. The brochure provided additional information about the study. A total of 30 individuals agreed to complete follow-up neuropsychological testing within a few months of the CSHA-3 test battery administration date.

Individuals involved in this study did not receive payment for their participation. Each CSHA participant provided informed consent at each wave of the study and ethical approval was independently granted from university ethics committees at all 18 CSHA study centres. Ethical approval for the current study was granted by the Ethics Committee of the University of Victoria. All participants who agreed to participate in the follow-up after CSHA-3 provided informed consent (see Appendix A) and were treated in accordance with the Tri-Council Policy Statement regarding human research (Medical Research Council of Canada, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada, 1998) and the standards delineated by the Canadian Psychological Association and the American Psychological Association. Thank-you letters (see Appendix A) were mailed to all participants subsequent to their completion of the study.

Materials

Neuropsychological data. The neuropsychological battery used in the CSHA included measures with available normative data for older adults (for a complete description, see Tuokko, Kristjansson, & Miller, 1995). The battery included the Information subtest from the Wechsler Memory Scale (WMS; Wechsler, 1975),

Buschke's Cued Recall paradigm (BCR; Buschke, 1984), the RAVLT (Lezak, 1995; Rey, 1964), the Benton Visual Retention Test (BVRT; Benton, 1974), an abbreviated form of the Token Test (Benton & Hamsher, 1978), measures of phonemic and semantic verbal fluency (FAS and Animals, respectively; Spreen & Strauss, 1998), the WAIS-R (Wechsler, 1981) Digit Symbol subtest, and short forms (Satz & Mogel, 1962) of the WAIS-R Similarities, Comprehension, and Block Design subtests. The neuropsychological battery used in CSHA-3 and the follow-up assessment included the same measures identified above with the exception of the RAVLT, BVRT, Token Test, and the WAIS-R Comprehension subtest. These 4 measures had been removed from the CSHA-3 protocol to abbreviate the assessment procedure and reduce the time required of participants.

A total of 10 neuropsychological measures were examined in Phase A of this study and 16 measures were examined in the remaining phases. Phase A included 5 measures of memory (WMS Information and 4 indices from the BCR), 1 measure of verbal abstraction (WAIS-R Similarities), 2 language measures (FAS and Animals), and 2 measures assessing visuospatial construction skill and psychomotor speed (WAIS-R Block Design and Digit Symbol subtests). The BCR indices were Trial 1 free recall (range = 0 - 12), Retrieval (total free recall over 3 learning trials; range = 0 - 36), Acquisition (total free and cued recall over 3 learning trials; range = 0 - 36), and Delayed free recall (range = 0 - 12). Phases B through G additionally included 4 more memory measures (BVRT and 3 indices from the RAVLT), 1 measure of social and practical judgment (WAIS-R Comprehension), and 1 measure of auditory language comprehension

(Token Test). The RAVLT indices were Trial 1 free recall (range = 0 - 15), Total (total free recall over 5 learning trials; range = 0 - 75), and Delayed free recall (range = 0 - 15).

Clinically significant change data. There were four indicators of clinically significant change used in this study. The first index, employed in Phases C and D, was the presence or absence of diagnostic change from CSHA-1 final diagnosis to CSHA-2 final diagnosis. The final diagnoses in the CSHA reflected the consensus diagnostic opinion of a multi-disciplinary team of mental health professionals based on all available information gathered during that specific wave of the study. In the absence of reliable neuropathological indicators for diagnosing dementia (see Bancher, Paulus, Paukner, & Jellinger, 1997; Mega et al., 1996; Nagy et al., 1998), the consensus diagnostic opinion based on clinical data is typically viewed as the "gold standard" for dementia diagnosis. Ivnik et al. (2000) used a similar standard in their examination of diagnostic accuracy. As previously noted, CSHA clinicians specifically identified persons with no cognitive impairment (NCI), cognitive impairment but not dementia (CIND), and dementia diagnosed according to DSM-III-R criteria (American Psychiatric Association, 1987). Wherever possible, subtypes of dementia were identified in the CSHA using National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer Disease and Related Disorders Association (NINCDS-ADRDA) criteria (McKhann et al., 1984) for Alzheimer Disease and ICD-10 criteria for vascular dementia and other forms of dementia (World Health Organization, 1987).

A second index of clinically significant change was a subjective report of change. CSHA participants provided responses to items on the Geriatric Depression Scale (GDS; Brink et al., 1982) during the screening stage including "Do you feel that you have more

problems with memory than most?" Responses to this question were coded as either "yes" or "no." This index served as the comparison standard for clinically significant change in Phase E. The third index of clinically significant change used in Phase F came from informant data that were obtained using the CAMDEX. Informants (usually a family member) were specifically asked whether the participant had exhibited a general decline in mental functioning and responses were coded as "yes" or "no." Finally, Phase G examined clinically significant change from the perspective of the consensus opinion of CSHA-2 clinicians regarding whether or not the participant had exhibited cognitive loss (not necessarily complete diagnostic change). Responses to this question were coded as either "yes" or "no." These ratings were made by clinicians at CSHA-2 who were blind to the assessment results from CSHA-1.

Data analyses

The main statistical analyses described below were completed with SPSS (SPSS Inc., 1998) and power analyses were conducted using the Power and Sample Size Calculation program (Dupont & Plummer, 1997). The distribution of variables and residuals (i.e., observed differences) for all test-retest scores were examined for normality by testing skewness and kurtosis. Given the large sample size involved in the CSHA-1 and CSHA-2 dataset, visual inspection of a histogram informed the final judgment regarding whether the residuals were deemed normally or non-normally distributed (Tabachnick & Fidell, 1996). All residuals from the prospectively collected data used in Phase A were normally distributed except scores on the WMS Information subtest and BCR Acquisition index. In the CSHA dataset, only the residuals for the BCR Acquisition data were non-normally distributed. Test-retest correlations used in the

calculation of reliable change scores were based on Pearson's correlation product moment correlation coefficients for normally distributed variables and Spearman's rho, a non-parametric measure of association, for all non-normally distributed variables in the study (Aron & Aron, 1994).

Group mean change for each neuropsychological measure in Phases A and B (Hypotheses 1 and 2) was examined using paired sample t-tests. In Phase D, group-level comparison of mean change in test performance among persons who remained NCI versus those who developed dementia (Hypothesis 3) was performed using Mann-Whitney U-tests. Bonferroni's correction for multiple comparisons was applied to adjust for the inflation of Type I errors in all group-level comparisons. In Phase A, which involved a total of 10 neuropsychological measures, the criterion for statistical significance was set at an alpha of 0.005 (i.e., 0.05 / 10). In Phases B and D, with 16 neuropsychological measures, the criterion for statistical significance was set at an alpha of 0.003 (i.e., 0.05 / 16).

Individual-level change scores were calculated for each participant on each measure using the formulae defined in Equations 2 through 12. Means and standard deviations for neuropsychological measures were based on available sample statistics. Phase A sample statistics (e.g., means, standard deviations, correlation coefficients) were calculated based on all 30 participants in the prospective portion of the study. Sample statistics used in Phases B through G came from the reference sample of 166 older adults who had been identified as NCI at CSHA-1 and CSHA-2 and who had completed neuropsychological testing at both waves. An estimate of the population mean for the

neuropsychological tests (required for the calculation of RCI_{SPEER}) was based on the mean neuropsychological test performance of all NCI persons at CSHA-1 (n = 575).

Guttman's reliability coefficients (required to calculate RC_{ID} and RC_{INDIV} in Phases B through G) could not be determined from the CSHA dataset given the manner in which these data had been coded. Alternatively, test-retest data from a subset of the 166 persons in the NCI reference sample were used to calculate Guttman's reliability coefficients at CSHA-1 and CSHA-2 (see Appendix B). This subset specifically included all CSHA participants from British Columbia (see Table 3) who were deemed NCI at CSHA-1 and CSHA-2 (n = 59). Data from these participants were recoded from the original test protocols. Guttman's reliability coefficients for Phase A were based on prospectively collected data. Given the speeded nature of the tests, Guttman's reliability coefficients could not be calculated for the WAIS-R Digit Symbol subtest and semantic verbal fluency (Animals), which in turn prevented the calculation of RC_{ID} for these measures. The RC_{INDIV}, as specified by Hageman and Arrindell (1999), was only calculated for r_{DD} values ≥ 0.40. The r_{DD} values that were employed in the calculation of RC_{ID} and RC_{INDIV} in Phase A and Phases B through G are presented in Appendix B.

Regression-based predictions of follow-up test scores were calculated using formulae listed in Appendix B. Equations were developed using the small sample of persons with NCI seen in Phase A (n = 30) and the larger reference sample for the remaining phases of the study (n = 166). Potential outliers were identified as cases with large standardized scores (i.e.., $z \geq 3.3$) on both the simple and multivariate analyses. No outliers were identified in the NCI sample in Phase A. A total of 13 individual test scores were removed from the analysis of NCI persons over the 5 year interval (Phases B

through G); this included 1 from the WMS Information subtest, 4 from BCR Acquisition, 4 from BCR Delayed Recall, 1 from WAIS-R Similarities, 2 from the Token Test, and 1 from the measure of phonemic fluency. These outliers were removed prior to the derivation of prediction equations (Tabachnick & Fidell, 1996). Baseline performance was the only entered predictor for SRB_{MCS} and SRB_{CH} change scores. For SRB_{MULT}, re-test scores were predicted based on initial test performance, age (years), education (years), and gender (1 = male; 2 = female).

To calculate reliable change in Phases A and B, alpha was set to a 0.10 level (two-tailed). Temkin et al. (1999) adopted a similar criterion for demarcating the boundary of "normal" change when comparing change score methods in their study. At this alpha level, reliable change values exceeding +1.645 indicated "reliable improvement" and values below -1.645 indicated "reliable deterioration." Values between ± 1.645 were deemed "unchanged." Classification accuracy was examined using a chi-square test to determine whether the proportion of cases classified as reliably changed by each method fit an expected pattern. It was expected that 5% of scores would be classified as "reliably deteriorated," 90% as "unchanged," and 5% as "reliably improved." Non-significant results on the chi-square test indicated that the observed pattern of classification did not deviate from the expected pattern (i.e., a positive finding). Chi-square tests could not be completed when one or more of the categories equaled zero (e.g., 0% improved and/or 0% deteriorated).

In Phases C through G, each index for clinically significant change - consensus diagnostic opinion, subjective ratings, informants' ratings, and clinicians' ratings - was coded as a dichotomous variable (i.e., dementia/cognitive loss versus no

dementia/cognitive loss). Classification ratings from the change scores were also dichotomized to reflect the presence or absence of statistically reliable deterioration in test scores. The criterion for reliable change was $z = -1.645$ (one-tailed) for all change scores in Phases C through G.

For each neuropsychological measure and each change score method, the sensitivity, specificity, positive predictive value, negative predictive value, and overall correct classification rate were calculated (Essex-Sorlie, 1995). These values were all expressed as percentages. In this study, sensitivity referred to the probability that a person who had declined to dementia (Phase D) or demonstrated cognitive loss (Phases E through G) had been correctly identified as showing reliable deterioration on the neuropsychological change scores. Specificity referred to the proportion of persons without dementia/loss who had been correctly identified as showing no reliable deterioration according to each specific change score methodology. The positive predictive value referred to the proportion of persons who actually evidenced clinically significant decline when the test results indicated that performance had reliably deteriorated. The negative predictive value referred to the proportion of individuals who were correctly identified as showing no clinically significant cognitive loss when the test results also indicated no decline. The overall correct classification accuracy was the percentage of persons with dementia/loss and no dementia/no loss who had been corrected classified according to their reliable test score changes.

To determine the statistical significance of the associations between reliable change and clinically significant indices of change, the dichotomous data in Phases D through G were analyzed using chi-square tests of significance with alpha set at a

conservative 0.01 level (Type I error rate = 1%). For each neuropsychological test and statistical method, an odds ratio (OR) was also calculated to provide a measure of the degree of association between reliable deterioration in test score performance and clinically significant change. An OR reflects the probability or likelihood of dementia/loss when there is evidence of reliable deterioration in test scores versus the incidence of dementia/loss when there is no reliable deterioration. Values near 1 indicate a lack of association and values over 3 indicate strong positive associations. ORs are frequently employed in epidemiological research and have been identified as a useful index in clinical neuropsychology (Bieliauskas et al., 1997; Sandercock, 1989). Ivnik et al. (2000) used ORs to compare the diagnostic accuracy among different diagnostic approaches.

A different set of analyses were conducted to examine the utility of each statistical method across several measures (in contrast to the OR, which was computed for each combination of neuropsychological test and statistical method). The sum of reliably changed scores from the entire battery of tests was calculated using each method and compared to the dichotomous indices of clinically significant change using non-parametric Spearman rho correlation coefficients. These sums and their relation to clinically significant indices of change were examined with regard to their receiver operating characteristics (ROC). ROC curves plot the trade off between sensitivity and specificity across a range of cutoff values (i.e., number of tests in the entire battery that show reliable deterioration using a specific change score method). The area under the curve (AUC) in ROC analyses provides a convenient way to compare the accuracy of each statistical method to a clinically significant standard of change. AUC values range

from 0.0 to 1.0; 0.50 is associated with random chance and an AUC value of 1.0 would indicate that a change score method is 100% accurate in discriminating between persons with dementia (or loss) and those without in the study sample. Larger AUC values indicate better overall performance and values that are 0.8 or greater are generally considered good.

To facilitate clinical use, cutoffs were identified for all change score methods to yield high sensitivity (i.e., ≥ 95%) and high specificity (i.e., ≥ 95%) relative to the "gold standard" of consensus diagnostic opinion. These cutoffs indicated the number of test scores in the CSHA battery that would need to evidence reliable deterioration in order to be confident that a diagnostic change from NCI to dementia had indeed occurred. Optimal cut-off values were also determined that yielded at least 90% sensitivity rates. This decision reflected an emphasis on detecting cognitive loss/dementia rather than accurately labeling non-dementia states, though in most clinical situations one would also wish to minimize the risk of diagnosing someone with dementia who does not have it.

Power analyses were calculated for the dichotomous data analyses planned in Phases D through G. The alpha level was set to 0.01. Given the sample size, power analyses indicated that the study design in Phase D would accurately detect odds ratios of 5.3 and greater with power at 0.80 (Type II error probability = 20%). Under similar parameters, Phases E, F, and G were expected to detect odds ratios greater than or equal to 5.1, 2.6, and 3.1, respectively. When power was set to 0.95 (Type II error probability = 5%) and alpha remained at 0.01, the study design would allow detection of odds ratios greater than 7.8 (Phase D), 7.4 (Phase E), 3.3 (Phase F), and 4.0 (Phase G).

Results

Following the preliminary analyses, the results are presented in order of the study hypotheses (1 through 15) with group-level findings preceding those related to the individual-level analyses.

Descriptive data for the CSHA-1 participants who were selected for inclusion in this study are presented in Table 1. Participants were predominantly Caucasian, ranged in age from 65 to 98, and had 10 years of education on average. Estimates of premorbid intelligence (Barona & Reynolds, 1984) ranged from low average to high average. The majority of participants were community-dwelling and independently functioning as reflected by high Activities of Daily Living (ADL) scores. The ADL scale ranged from 0 to 28 with lower scores indicating greater need of assistance. This measure was based on an informant report of the participant's abilities in 14 areas including general self-care, mobility, phone use, meal preparation, housework, shopping, ability to manage finances and medication management. Follow-up data, including the proportion of persons who did not participate in CSHA-2, are listed in Table 2.

The NCI participants from CSHA-1 who participated in CSHA-2 were younger ($M = 78.3$ years, $SD = 6.3$) than those who did not participate due to death, refusal, or being lost to follow-up ($M = 80.8$ years, $SD = 7.0$), $F(1, 573) = 21.1$, $p < .001$. Individuals who participated in both CSHA-1 and CSHA-2 were more educated ($M = 10.6$ years, $SD = 3.8$) than those who were not seen at follow-up ($M = 9.8$ years, $SD = 3.8$), $F(1, 570) = 6.6$, $p < .01$. The latter groups, however, did not differ with respect to gender, $X^2(1, N = 575) = 0.10$, ns.

Participants who were judged to have NCI at both CSHA-1 and CSHA-2 were compared with respect to whether or not they completed the neuropsychological component at CSHA-2. Persons who completed the assessment did not differ from those who did not in terms of age, $F(1, 206) = 0.00$, ns, education, $F(1, 205) = 0.76$, ns, or gender, $X^2(1, N = 208) = 0.69$, ns. The groups also did not differ on any of the neuropsychological measures at baseline (p's > 0.05) with one exception. Individuals who completed the assessment at CSHA-2 had higher scores on the WAIS-R Digit Symbol subtest at baseline than those who did not complete the assessment at follow-up, $F(1, 196) = 8.80$, $p < 0.01$.

NCI participants who developed dementia at follow-up were also examined to determine whether differences existed between those who did and did not complete the neuropsychological component at CSHA-2. These two groups did not differ in terms of age, $F(1, 47) = 0.85$, ns, education, $F(1, 47) = 0.59$, ns, gender, $X^2(1, N = 49) = 0.00$, ns, or any of the neuropsychological measures at baseline (p's > 0.05) with the exception of performance on the WAIS-R Comprehension subtest. Individuals who did not complete the assessment at CSHA-2 had lower scores on this measure at baseline relative to those who had participated in the neuropsychological component, $F(1, 47) = 5.98$, $p < 0.05$.

Table 3 contains descriptive data for the subset of CSHA-1 NCI participants who remained NCI at CSHA-2 and completed the neuropsychological battery at both waves of the study. This group of 166 comprised the reference sample used to calculate most of the sample statistics used in the reliable change calculations. These individuals averaged 77 years of age and had a mean 11 years of education. There were slightly more females (57%) than males in the sample, which was consistent with Canadian population

demographics for persons over age 64 (Statistics Canada, 2002). The majority of persons in this reference sample were urban community-dwelling, functionally capable individuals who reported having at least good physical health. The mean test-retest interval for these participants was 58 months (approximately 4.8 years) and ranged from 42 to 70 months. Test-retest correlations over this interval were highly significant (all p's < 0.001) for all but the BCR Acquisition index, which showed no relation between baseline and follow-up performance, $r = 0.03$, ns. Of the remaining measures, correlations ranged from low (WMS Information, $r = 0.33$, $p < 0.001$) to high (WAIS-R Digit Symbol subtest, $r = 0.85$, $p < 0.001$) with most measures yielding test-retest correlation coefficients in the moderate range.

Data pertaining to the CSHA-1 NCI group that progressed to dementia at follow-up and completed neuropsychological testing at both waves are presented in Table 4. This group of 20 participants ranged in age from 67 to 89 years and had a mean 10 years of education. They demonstrated, on average, a decline of 15 points on the 3MS and 5 points on the MMSE over the 5-year test-retest interval. Like the NCI reference sample, the mean retest interval for this group was also 58 months. The retest interval ranged from 52 to 68 months. At follow-up, this group of participants included 9 diagnosed with probable Alzheimer Disease, 4 with possible Alzheimer Disease, 5 with vascular dementia, 1 with dementia due to Parkinson's Disease, and 1 person with an unclassified dementia.

Data from the sample of 30 CSHA-3 participants with NCI who were re-evaluated as part of the prospective portion of the study are presented in Table 5. Participants ranged in age from 75 to 97 and had a mean of 12 years of education. The

average test-retest interval for these participants was 3 months (ranging from 1 to 7 months). Neuropsychological summary data for this sample, including mean performance at baseline, mean performance at follow-up, and mean difference scores, are presented in Table 6. In spite of the relatively small sample size, test-retest correlations were all statistically significant (p's < 0.05). The strength of the correlations ranged from low (BCR Delayed recall, $r = 0.36$, $p < 0.05$) to high (FAS, $r = 0.89$, $p < 0.001$). There was no variability in performance on the BCR Acquisition score at follow-up (all participants scored perfectly) and therefore this measure was dropped from most Phase A analyses.

Group-level analyses of change

Hypothesis 1. According to this hypothesis, practice effects were expected among cognitively intact older adults from Phase A (i.e., the relatively short interval spanning a few months). As shown in Table 6, these NCI participants evidenced higher scores on all neuropsychological measures with the exception of a slight decline in performance on the WAIS-R Similarities subtest. Improvement on the BCR was close to reaching statistical significance on two indices; Trial 1 performance, $t(29) = 2.80$, $p = 0.009$, and the Retrieval score, $t(29) = 2.87$, $p = 0.008$. None of the test-retest differences, however, were significantly improved when assessed using t-tests with a Bonferroni correction (all p's > 0.005). Therefore, contrary to Hypothesis 1, the NCI group that was followed over a few months did not evidence substantive practice effects on any of the measures that were administered. Subsequent correlational analyses revealed that individual difference scores (i.e., practice effects) were not associated with age for any measure (all p's > 0.05).

Hypothesis 2. It was predicted that NCI participants who were followed over 5 years would show either stable or deteriorated mean test performance. The test findings presented in Table 7 show that CSHA-1 NCI participants scored lower, on average, on many neuropsychological measures when they were retested at CSHA-2. There were no significant practice effects or trends in this direction with the exception of a slight improvement on the WAIS-R Similarities subtest. In contrast, statistically significant declines in test-retest performance were found on speeded measures including the WAIS-R Block Design subtest, $t(159) = 5.31$, $p < 0.001$, and the WAIS-R Digit Symbol subtest, $t(152) = 8.59$, $p < 0.001$. Significant deterioration was also evident on untimed measures including the WMS Information subtest, BCR Trial 1 and Retrieval, and the RAVLT Total and Delayed recall conditions (all p's < 0.003). Test-retest change on the BCR Delayed recall approached a level of statistical significance using a Bonferroni correction but did not exceed it, $t(161) = 2.96$, $p = 0.0035$. No other measures showed a significant deterioration over the 5-year interval (p's > 0.003). These findings of stable and deteriorated test performances over the long test-retest span are consistent with the study hypothesis.

Subsequent correlational analyses revealed weak, but statistically significant, associations between age and difference scores (i.e., $X_2 - X_1$) on the WMS Information subtest, $r = -0.161$, $p < 0.05$; BCR Retrieval, $r = -0.184$, $p < 0.05$; BCR Acquisition, $r = -0.159$, $p < .0.05$; BCR Delayed Recall, $r = -0.178$, $p < 0.05$; semantic fluency (Animals), $r = -0.158$, $p < 0.05$; and the WAIS-R Digit Symbol subtest, $r = -0.279$, $p < 0.001$.

Hypothesis 3. According to this hypothesis, persons who progressed to dementia by follow-up were expected to demonstrate global decline across all cognitive test

measures. The hypothesis was tested by comparing the sample of 166 older adults who were NCI at both CSHA-1 and CSHA-2 with the sample of 20 persons with NCI at CSHA-1 and dementia at CSHA-2. Group-level analyses revealed that persons identified with dementia at follow-up evidenced more decline, on average, than persons who remained NCI on most of the CSHA neuropsychological measures. Using Mann-Whitley U-tests, significantly more decline was found among persons who developed dementia on the WMS Information subtest, $U = 485.0$, BCR Trial 1, $U = 497.5$, BCR Retrieval, $U = 402.5$, BCR Acquisition, $U = 290.5$, BCR Delay recall, $U = 259.5$, RAVLT Total, $U = 498.0$, BVRT, $U = 494.5$, WAIS-R Similarities subtest, $U = 825.0$, WAIS-R Comprehension subtest, $U = 652.5$, Token Test, $U = 498.0$, Letter fluency, $U = 659.5$, and Category fluency, $U = 646.5$ (all p's < 0.001). A similar trend existed on the remaining neuropsychological measures but these did not reach statistical significance using a Bonferroni correction (i.e., $p < 0.003$); RAVLT Trial 1, $U = 760.0$, $p = 0.004$, RAVLT Delayed recall, $U = 690.5$, $p = 0.018$, WAIS-R Block Design, $U = 1053.5$, $p = 0.028$, and WAIS-R Digit Symbol, $U = 513.0$, $p = 0.004$. Overall, these findings provided partial support for Hypothesis 3 with findings of statistically significant decrements in most measures of memory, language, abstraction, and judgment in conjunction with statistically non-significant trends in the expected direction on the remaining measures of memory, visuoconstructional skill, and psychomotor speed.

Individual-level analyses of change

Appendix C contains all data pertaining to the classification accuracy for each measure and statistical method used in Phases A, B, and C, including a chi-square goodness-of-fit analysis. Classification accuracy data for Phases D through G, which are

based on dichotomous variables, are presented in Appendix D; this appendix includes the results of a chi-square test of significance (Chi^2), sensitivity (SENS), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV), overall classification accuracy (CA), the odds ratio (OR), and Spearman's rho (r). Classification accuracy data, as shown in the appendices, could not be computed for all measures and methods. As previously noted, Guttman's reliability coefficients could not be calculated for 2 speeded tests in the CSHA test battery which precluded the computation of RC_{ID} for these measures. Several of the r_{DD} values in this study were lower than the 0.40 criterion specified by Hageman and Arindell (1999) which meant that the RC_{INDIV} classification as improved, unchanged, or deteriorated could not be calculated for any of the measures in Phase A except BCR Retrieval. In Phases B through G, low r_{DD} values precluded the calculation of RC_{INDIV} for BCR Trial 1 and Delayed recall, RAVLT Trial 1, BVRT, WAIS-R Comprehension, Block Design, and Digit Symbol subtests, phonemic fluency (FAS), and semantic fluency (Animals).

Hypothesis 4. It was expected that certain statistical methods, such as the SRB_{CH} and SRB_{MULT}, would be more accurate than other change score methods in classifying persons as showing improvement, deterioration, and no change. The SD method was expected to be the least accurate method. The findings from Phase A demonstrated significant overlap in classification accuracy of normal change among the statistical methods. Overall, the hypothesized pattern of superiority of certain statistical methods over others was not supported by the CSHA data.

As shown in Table 8, the RCI_{HSU} yielded the best classification accuracy rates with 7 out of 10 measures conforming to the expected classification distribution in the

chi-square goodness-of-fit analyses. The RCI_{JT} and RCI_{CHEL} (using formulae a and c) were the next most accurate set of change score methods followed by $RCI_{JT}b$, RCI, the 3 regression-based methods, RCI_{SPEER} and $RCI_{CHEL}b$. The SD method accurately classified performance on 2 of the 10 measures and the RC_{INDIV} and RC_{ID} methods resulted in no correct classifications. Notably, the RC_{INDIV} failed to result in any correct classifications even when the analyses were repeated using measures associated with low r_{DD}s.

Contrary to prediction, the RCI_{HSU}, RCI_{JT}, and RCI_{CHEL} resulted in better overall classification accuracy than the more sophisticated SRB methods in this sample over a relatively short span of time. The RCI_{JT} and RCI_{CHEL} yielded comparable results using formulae a and c, but accuracy rates were reliably weaker using formula b where the SED was calculated using separate SEMs from time 1 and time 2. The 3 regression-based methods performed comparably to each other, and overlapped considerably with the above mentioned RCIs, but often classified too few individuals as showing reliable improvement in this sample. The opposite pattern was seen with the RCI_{SPEER}, and to a slightly lesser extent, the SD and RCI methods, which tended to classify too many individuals as showing reliable improvement in Phase A. The $RCI_{JT}b$ and $RCI_{CHEL}b$ classified fewer people as unchanged than the other methods with errors in both directions of improvement and deterioration.

Hypothesis 5. According to this hypothesis, the regression-based methods were expected to yield narrower prediction intervals than the RCI methods, which in turn would be narrower than those associated with the SD method. The Phase A data pertaining to the width of the prediction intervals are presented in Table 9. On the one measure for which RC_{INDIV} could be appropriately calculated (i.e., $r_{DD} \geq 0.40$), it

generated the narrowest prediction interval. This pattern consistently held when the RC_{INDIV} was calculated for the remainder of measures with r_{DD}s below 0.40. The next narrowest intervals were found using the RCI_{SPEER} and RCI methods, both of which employed the SEM. In comparison, the SD method, which set the boundary of normal change within ± 1SD, resulted in a slightly larger interval width. Next, the RCI_{HSU} yielded prediction intervals that were greater than ± 1SD for most measures but were generally smaller than those associated with the two regression-based methods across the neuropsychological measures. The RCI_{JT}, RCI_{CHEL}, and RC_{ID} consistently yielded the largest prediction intervals. As seen in Table 9, the results from Phase A were not consistent with the expected pattern delineated in Hypothesis 5.

Hypothesis 6. No statistical method was expected to show a pattern of reliable deterioration on measures of memory and one (or more) other cognitive domains at a rate greater than chance. In Phase A, which employed a total of 10 measures, cognitively healthy participants had a 50% probability of having one or more reliably deteriorated test scores by chance alone (0.05 x 10 measures). The RCI_{CHEL}b, at 66.7% classified slightly more than an expected proportion of individuals as having one or more reliably deteriorated sets of scores. Most other methods (see Table 8) classified between 30% to 50% of the sample as having one or more reliably deteriorated test scores. In contrast, the RC_{ID} and RC_{INDIV} were very conservative and almost consistently classified individuals as showing no reliable change in any test performance at rates far below chance levels (i.e., 0% and 3%, respectively).

In a neuropsychological test battery with 5 memory measures and 5 measures assessing other cognitive domains, the combined likelihood of showing reliable

deterioration in memory and one (or more) other cognitive domains is about 6% (0.05 x 5 measures x 0.05 x 5 measures). As shown in Table 8, none of the statistical methods misclassified individuals as showing reliable decrements in multiple areas of cognition at a rate greater than chance. These findings from Phase A supported Hypothesis 6.

Hypothesis 7. Using Phase B data, it was expected that certain statistical methods, such as the SRB_{CH} and SRB_{MULT}, would be more accurate in classifying cognitively healthy persons as showing improvement, deterioration, or no change than others. There were distinct differences in accurately classifying normal change among the various methods. Contrary to expectation, the $RCI_{CHEL}c$ resulted in the highest rate of classification accuracy with all 16 measures conforming to the expected distribution (see Table 10). This was followed closely, in terms of accuracy, by $RCI_{CHEL}a$ and then the 3 SRB methods. Among the latter, the models based on simple regression (i.e., SRB_{MCS} and SRB_{CH}) showed a slight advantage over the multiple regression model. The RCI_{HSU} and RCI_{JT} (formulae a and c) were moderately accurate with correct classifications on 11 of the 16 neuropsychological measures shown by the chi-square analyses. Notably, the RCI_{HSU} was generally comparable in accuracy to the regression-based methods across most neuropsychological measures (see Appendix C). The RCI_{JT} (formula a and c), as noted above, was reasonably accurate. The Phase B data consistently indicated a superiority of using formulae c and a in calculating the SED for RCI_{CHEL} and RCI_{JT} relative to formula b.

In decreasing order of accuracy, low to poor classification accuracy rates were found using the $RCI_{CHEL}b$, $RCI_{JT}b$, RCI, SD, RC_{INDIV}, RC_{ID}, and RCI_{SPEER} methods. These statistical methods were less consistent, or simply inadequate, in classifying

normal change in this sample of cognitively healthy adults. On several measures, the RC_{ID} was too conservative and classified persons as unchanged; on other measures, it tended to classify too many persons as deteriorated. An increased risk of misclassification as both reliably improved and reliably deteriorated was found using the SD method, RCI, RCI_{SPEER}, $RCI_{JT}b$, and $RCI_{CHEL}b$ methods. The $RCI_{CHEL}b$ was particularly prone to misclassify individuals as showing reliable improvement and the $RCI_{JT}b$, most notably, along with the SD and RCI methods tended to misclassify individuals as exhibiting reliable test score deterioration. The RC_{INDIV} could not be calculated for the majority of neuropsychological measures but among the remaining measures, it tended to classify too many individuals as showing reliable deterioration. Overall, the expected pattern of superiority among change score methods was not supported by the Phase B data.

Hypothesis 8. Prediction intervals associated with each statistical method from Phase B are shown in Table 11. The RC_{INDIV}, on average, yielded the narrowest prediction intervals for measures with values of $r_{DD} \geq 0.40$ (as well as for measures with $r_{DD} < 0.40$). In order of increasing width, the interval size grew with the SD method, RCI_{SPEER}, and RC_{ID}. The $RCI_{JT}b$, $RCI_{CHEL}b$, SRB_{MULT}, SRB_{MCS}, and RCI_{HSU} had moderate size prediction intervals. In contrast, the RCI_{JT} and RCI_{CHEL} (formulae a and c) were associated with the widest prediction intervals. The results from Phase B were not consistent with the expected pattern of results delineated in Hypothesis 8.

Hypothesis 9. No statistical method was expected to show a pattern of reliable deterioration on measures of memory and one (or more) other cognitive domains at a rate greater than chance. In Phase B, one would expect a misclassification rate of about 15%

by chance alone using a neuropsychological test battery comprised of 9 memory measures and 7 measures assessing other cognitive abilities (0.05 x 9 measures x 0.05 x 7 measures). The RC_{INDIV} and $RCI_{CHEL}c$ were the change score methods that were least likely to categorize a cognitively intact sample of older adults as showing reliable deterioration across multiple areas of cognition (i.e., memory plus at least one other area). As seen in Table 10, these methods misclassified approximately 7% of the sample of showing declines in a pattern consistent with a dementia. The $RCI_{CHEL}a$, RC_{ID}, RC_{HSU}, and the 3 regression-based methods were also associated with lower than chance rates of misclassification across multiples areas of decline (all less than 10% of the sample misclassified). In contrast, the RCI_{SPEER} and RCI_{JT} (formulae a and c) were associated with greater than chance levels of misclassification ranging from 19% to 25%. The $RCI_{CHEL}b$, RCI, and SD methods were notably poor in this regard with 32% to 42% of the sample showing reliable deterioration in memory and at least one other cognitive domain. Among all of the statistical methods, the $RCI_{JT}b$ was the least accurate in this phase. The use of the $RCI_{JT}b$ inappropriately classified more test scores as reliably deteriorated than any other method and incorrectly identified the greatest percentage of the cognitively healthy adult sample, at nearly 50%, as having cognitive impairments consistent with a dementia (see Table 10). Given the low accuracy rates associated with some of the statistical methods, the results of Phase B did not support Hypothesis 9.

Hypothesis 10. According to this hypothesis, all NCI individuals at baseline who developed dementia by follow-up were expected to show a pattern of reliable decline on a test(s) of memory and a test(s) in one other cognitive domain using each of the change score methods. Data from Phase C involving a small sample of 20 individuals who

progressed to dementia over a 5-year interval were examined to address this issue. The findings from this phase are summarized in Table 12.

Among the methods with relatively high accuracy rates, the $RCI_{JT}b$ was the most accurate in this phase. The $RCI_{JT}b$ detected reliable change in two or more areas of cognition for each and every one of the participants in this small sample. The RCI, RCI_{SPEER}, $RCI_{CHEL}b$ and SD method were next best followed closely by the 3 SRB methods, RCI_{HSU}, and the RCI_{JT} (formulae a and c). The $RCI_{CHEL}a$ and $RCI_{CHEL}c$ were relatively less accurate than the other methods but still correctly identified 80% of the sample as having significantly lower test scores.

From the perspective of classifying reliable change in multiple areas of cognition, the RC_{INDIV} and RC_{ID} did not fare as well as the other methods with only 60% and 70% accuracy rates, respectively. The RC_{ID}, in particular, was poor as there were two individuals who were deemed to have developed dementia by consensus opinion who did not show reliable deterioration on any of the 16 neuropsychological measures using this statistical method. Overall, nearly all of the statistical methods, with the exception of RC_{INDIV} and RC_{ID}, were good or adequate in classifying change in NCI persons who had progressed to a dementia with accuracy rates ranging from 80% to 100%. These findings provide partial support for the study hypothesis.

Hypothesis 11. Hypothesis 11 pertained to the diagnostic classification accuracy of the various statistical methods using Phase D data, with the greatest accuracy expected using the SRB_{CH} and SRB_{MULT} methods, followed by SRB_{MCS}, RC_{INDIV}, RC_{ID}, RCI_{CHEL}, RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and the SD method for each measure. A total of 186 persons were examined who had no cognitive impairment at baseline; 166 of whom

showed no diagnostic change over time and 20 persons who had progressed to dementia by follow-up. As summarized below, the hypothesized pattern of superiority among statistical methods was not supported using any of the CSHA neuropsychological measures (see Appendix D for details).

Chi-square analyses revealed that the WMS Information subtest, all 4 memory indices from the BCR, BVRT, WAIS-R Comprehension, Token Test, and phonemic fluency (FAS) were significantly associated (all p's < 0.001) with diagnostic change across all statistical change score methods. Using these same analyses, the only statistical methods that were consistently associated with diagnostic change across all CSHA neuropsychological measures were the RCI_{HSU}, RCI_{SPEER}, and SRB_{MULT} (all p's < 0.001). More detailed results regarding the sensitivity, specificity, and odds ratios associated with each measure/method are provided below.

Test score change on the WMS Information subtest resulted in relatively high sensitivity and specificity rates with classification accuracy ranging from 62% to 92% depending upon the statistical method. All ORs were extremely large with most scattered in the 20 to 50 range. The RCI_{ID} yielded the highest odds ratio at 70.7 [95% CI = 7.9, 629.4] but was associated with very low sensitivity (30%). SRB_{MCS} and SRB_{CH} had the next highest odds ratios (OR = 51.3 [95% CI = 14.8, 177.9]) followed by the RCI_{HSU} (OR = 43.8 [95% CI = 11.7, 163.6]). The $RCI_{CHEL}c$, $RCI_{CHEL}b$, and $RCI_{JT}b$ were also highly accurate and comparable to one another (ORs > 30); these methods were followed by the SRB_{MULT}, RC_{INDIV}, $RCI_{CHEL}a$, $RCI_{JT}c$, $RCI_{JT}a$, RCI_{SPEER}, and RCI (ORs > 20). Among all of the change scores, the SD method had the weakest association with diagnostic change on the WMS Information subtest (OR = 8.6 [95% CI = 2.4, 30.4]).

Test score differences on all BCR indices also demonstrated a strong and significant relation with diagnostic change. Reliable change on BCR Trial 1 recall was most strongly associated to diagnostic change using the SRB_{MULT} (OR = 77.0 [95% CI = 20.8, 284.2]) and RCI_{HSU} (OR = 64.9 [95% CI = 16.8, 251.0]). The RCI_{CHEL}a, b, and c were the next strongest statistical methods followed by the remaining two regression-based methods (all ORs > 40). RCI_{SPEER} yielded an odds ratio of 32.6 [95% CI = 8.8, 119.7] followed by the RCI_{JT}a, b, and c methods (OR = 21.3 [95% CI = 7.0, 65.0]). The RCI and SD methods were the weakest associated with this measure (OR = 11.8 [95% CI = 3.7, 37.3]. Odds ratios could not be calculated for RC_{INDIV} and RC_{ID}.

Test score change on the BCR Retrieval was strongly related to diagnostic accuracy using all change score methods as shown by the very large odds ratios. The methods with the highest odds ratios were RCI_{HSU} (OR = 68.2 [95% CI = 14.4, 322.6]) and the SRB methods (ORs ranging from 57.3 to 62.5). RCI_{JT}c yielded the next strongest association followed by RCI_{CHEL}a and c (OR = 40.1 [95% CI = 11.5, 139.7]), RCI_{SPEER}, RCI_{JT}a, RCI_{CHEL}b, RC_{ID} and RC_{INDIV}. The RCI_{JT}b, RCI, and the SD method had the weakest association (OR = 16.3 [95% CI = 4.5, 59.4]) with this memory measure among all of the change score methods.

The BCR Acquisition score, though non-normally distributed, was also strongly related to diagnostic accuracy. The RCI, RCI_{JT}a and c, RCI_{CHEL}a, and RC_{ID} were all associated with extremely large odds ratios (OR = 137.8 [95% CI = 29.6, 642.1]) as were the RCI_{HSU} and RCI_{SPEER} methods (OR = 102.7 [95% CI = 24.5, 429.7]). Relatively smaller odds ratios were found using the RCI_{JT}b, RCI_{CHEL}b and c methods followed by

RC_{INDIV} and SD methods. On this measure, the 3 SRB methods had the lowest diagnostic accuracy though were still associated with very large odds ratios (all ORs > 30).

Overall classification accuracy rates on the BCR Delayed Recall measure varied by method and ranged from 82% to 91%. Reliable change on this memory index was strongly associated with diagnostic change across all methods (all ORs > 30). The SRB_{CH}, SRB_{MCS}, RCI_{SPEER} were most accurate (OR = 64.0 [95% CI = 13.4, 301.4]) followed closely by the RCI_{HSU} (OR = 62.5 [95% CI = 15.8, 246.4]) and SRB_{MULT} (OR = 42.6 [95% CI = 11.2, 162.5]). The SD and RCI method had the weakest associations (OR = 32.5 [95% CI = 7.7, 161.3]) and there were no differences among the remaining methods in terms of diagnostic accuracy. ORs could not be computed for RC_{ID} and RC_{INDIV}.

In comparison to the BCR, the RAVLT yielded considerably lower accuracy rates. Odds ratios on Trial 1 recall ranged from 16.1 (RCI_{HSU} [95% CI = 4.2, 61.4]) and 14.6 (SRB_{CH} and SRB_{MCS} [95% CI = 4.3, 49.9]) down to 3.8 (SD method [95% CI = 1.4, 10.6]). SRB_{MULT} and RCI_{SPEER} were comparable and more accurate than the $RCI_{CHEL}b$, $RCI_{JT}b$, and RCI methods. The $RCI_{JT}a$, $RCI_{JT}c$, $RCI_{CHEL}a$, and $RCI_{CHEL}c$ yielded relatively weak associations (OR = 4.5 [95% CI = 1.0, 19.2]) using this memory index.

Diagnostic accuracy improved slightly using the RAVLT total score. Two of the regression-based methods (SRB_{CH} and SRB_{MCS}) and the RCI_{HSU} resulted in large associations (OR = 53.3 [95% CI = 12.7, 223.3]), followed by SRB_{MULT}. The next best group of methods included the RCI_{SPEER}, $RCI_{CHEL}a$ and c, and $RCI_{JT}a$ and c. On this memory measure, the RCI and SD methods outperformed the $RCI_{CHEL}b$, $RCI_{JT}b$, RC_{INDIV}, and the RC_{ID} (OR = 3.5 [95% CI = 1.1, 10.8]).

Delayed recall on the RAVLT had the strongest association with diagnostic accuracy using the SRB_{MULT} (OR = 15.6 [95% CI = 4.2, 58.1]) followed by the SRB_{CH} and SRB_{MCS}, RCI_{SPEER}, RCI_{HSU}, and RCI_{CHEL}a and c. Relatively weaker associations were found using the remaining methods (ORs ranging from 3.0 to 4.5).

Classification accuracy using change scores on the BVRT was quite high and ranged from 85% to 93%. ORs were statistically significant with most in the 10 to 25 range. The weakest associations (OR = 10.3 [95% CI 3.5, 30.3]) were found using the RCI_{JT}b, RCI, and SD methods. Accuracy improved with the RCI_{HSU} and RCI_{SPEER} (OR = 18.4 [95% CI = 5.7, 59.0]) and increased further with the 3 SRB methods, RCI_{JT}a and c, and RCI_{CHEL}b and c. The highest degree of accuracy on the BVRT was found with the RCI_{CHEL}a (OR = 45.3 [95% CI = 10.2, 200.6]). Odds ratios could not be computed for RC_{ID} and RC_{INDIV}.

Reliable change on the WAIS-R Similarities subtest was specific but lacked sensitivity to dementia. Most ORs ranged from 4 to 8 using this measure. The RC_{ID} had the highest association with diagnostic accuracy using the Similarities subtest (OR = 21.7 [95% CI = 3.7, 128.6]) and most other methods showed considerably weaker relations. The RCI_{HSU} yielded an odds ratio of 9.3 [95% CI = 3.1, 27.4] with declining strength of association found using the SRB_{CH}, SRB_{MCS}, RCI_{SPEER}, and SRB_{MULT}. The remaining methods were comparable with relatively weak relations to diagnostic accuracy (ORs ranging from 4.6 to 5.3).

There was little variability in classification accuracy on the WAIS-R Comprehension subtest across statistical methods (range from 87% to 90%). Change on this measure was consistently and fairly strongly related to diagnostic change (ORs

ranging from 10 to 20). The SRB_{CH} alone had the strongest association (OR = 20.3 [95% CI = 6.2, 65.9]) followed by $RCI_{JT}a$, $RCI_{JT}c$, $RCI_{CHEL}a$ and $RCI_{CHEL}c$ (OR = 17.7 [95% CI = 5.6, 55.6]), RCI_{HSU}, SRB_{MCS}, SRB_{MULT}, and RCI_{SPEER}. $RCI_{JT}b$, $RCI_{CHEL}b$, RCI, and the SD method yielded the weakest associations (OR = 10.5 [95% CI = 3.6, 30.5]) with this measure. Odds ratios could not be calculated for RC_{ID} and RC_{INDIV}.

Like the WAIS-R Comprehension subtest, change scores on the Token Test were consistently related to change in diagnostic status (with most ORs approximately equal to 10). Change in test performance on the Token Test was most accurately classified using the $RCI_{CHEL}a$ (OR = 30.1 [95% CI 6.9, 131.7]) followed by the RCI_{HSU} (OR = 19.4 [95% CI = 6.0, 63.0]), SRB_{CH} and SRB_{MCS} (OR = 13.5 [95% CI = 4.6, 39.6]). There was little difference among the remaining statistical methods in classification accuracy (ORs ranging from 10.6 to 12.6).

On a measure of phonemic fluency (FAS), classification accuracy ranged from 73% to 90%. ORs ranged from 5 to 18. The $RCI_{CHEL}a$ resulted in the highest association (OR = 18.8 [95% CI = 4.0, 87.1]) followed by RCI_{HSU} and RCI_{SPEER} (OR = 14.4 [95% CI = 3.9, 53.6]). The next strongest group of methods included the SRB_{MULT}, RC_{ID}, $RCI_{CHEL}c$, and $RCI_{JT}a$ and c. The strength of the association decreased with $RCI_{JT}b$, the remaining regression-based methods, SD, and $RCI_{CHEL}b$. For this verbal fluency measure, the weakest association with diagnostic accuracy was found using the RCI method (OR = 5.4 [95% CI = 2.0, 14.7]). The odds ratio could not be calculated using the RC_{INDIV}.

On the semantic word fluency task, reliable change defined using the SRB_{MULT} method was strongly associated with diagnostic classification (OR = 32.0 [95% CI = 7.0,

147.2]) followed by SRB_{CH}, SRB_{MCS}, and RCI_{HSU} (OR = 24.6 [95% CI = 6.5, 92.9]). The RCI_{SPEER} method (OR = 16.2 [95% CI = 5.1, 51.1]) resulted in greater accuracy than $RCI_{JT}b$ (OR = 6.1 [95% CI = 2.1, 17.9]) but all other methods were comparable with relatively weak associations to diagnostic change (ORs ranging from 3.8 to 3.9). Odds ratios could not be calculated using the RC_{ID} and RC_{INDIV}.

Classification accuracy based on test score changes on the WAIS-R Block Design subtest ranged from 64% to 89%. RCI_{SPEER} yielded the highest odds ratio on this measure (OR = 16.6 [95% CI = 5.7, 48.5]) followed by RCI_{HSU} (OR = 11.1 [95% CI = 2.8, 42.9]) and SRB_{MULT} (OR = 9.2 [95% CI = 2.4, 33.9]). Accuracy decreased with SRB_{CH}, SRB_{MCS}, $RCI_{JT}a$, $RCI_{CHEL}c$, $RCI_{JT}c$, $RCI_{CHEL}a$, $RCI_{CHEL}b$, RCI, and $RCI_{JT}b$. Diagnostic accuracy was particularly low using the SD method (OR = 1.8 [95% CI = 0.6, 4.8]). Odds ratios could not be calculated using the RC_{ID} and RC_{INDIV}.

Reliable change calculated using the $RCI_{CHEL}c$ was most diagnostically accurate using the WAIS-R Digit Symbol subtest (OR = 31.9 [95% CI = 7.3, 139.5]). The next best methods included the SRB_{MULT} and RCI_{HSU} (ORs > 20) followed by the SRB_{CH}, SRB_{MCS}, $RCI_{CHEL}a$, and SD method (ORs > 10). Weaker associations were found using the RCI_{SPEER}, $RCI_{JT}a$ and c, $RCI_{CHEL}b$, and RCI. The $RCI_{JT}b$ method had the weakest association with diagnostic accuracy (OR = 2.1 [95% CI = 0.6, 7.1]) using this measure. Odds ratios could not be calculated using the RC_{ID} and RC_{INDIV}.

In summary, the individual analyses from this phase did not support the study hypothesis. Though the SRB methods were often accurate, they were not consistently more accurate than the RCI methods (e.g., RCI_{CHEL} and RCI_{HSU}) and some of the more sophisticated RCI methods (e.g., RC_{ID} and RC_{INDIV}) did not outperform less sophisticated

change score models. The findings also indicate that changes on the BCR and WMS Information subtest were most strongly and consistently associated with diagnostic change. The BVRT, WAIS-R Comprehension, Token Test, and phonemic fluency (FAS) also showed significant, but slightly weaker, associations with change in diagnostic status. Changes on the RAVLT, semantic fluency, and the WAIS-R Block Design and Digit Symbol subtests were also diagnostically useful but only with specific change score methods such as the SRB methods, RCI_{HSU}, or RCI_{CHEL}. Amidst all of the combinations of change score methodology and neuropsychological measure, only the WMS Information subtest (using RCI_{HSU}), BCR Trial 1 (using RCI_{HSU} or RCI_{SPEER}), BCR Total Retrieval (using RCI_{HSU}), and the BCR Delay Recall (using RCI_{SPEER}, SRB_{MCS}, or SRB_{CH}) were able to accurately classify individuals with 85% sensitivity and 85% specificity into categories of no dementia versus dementia.

Hypothesis 12. This hypothesis addressed the diagnostic accuracy of the sum of reliable changes across all 16 measures in the neuropsychological battery using each statistical method. Data from Phase D were employed to examine the relation between the sum of reliable changed test scores and diagnostic change as determined by a consensus opinion of health professionals. The findings are summarized in Table 13. All change score methods produced significant correlations between the sum of reliable changes and consensus diagnosis (all p's < 0.001). The strength of the correlations were largely consistent among the statistical methods (range = -0.54 to -0.43) but subtle differences were found. The RCI_{HSU} yielded the strongest correlation (Spearman's rho = -0.545) followed by the RCI_{CHEL}a and SRB_{CH} methods. The remaining methods, in decreasing order of strength of association with diagnostic change, were SRB_{MCS},

$RCI_{CHEL}c$, SRB_{MULT}, RCI_{SPEER}, $RCI_{JT}a$, $RCI_{JT}c$, $RCI_{CHEL}b$, RCI, $RCI_{JT}b$, SD, RC_{INDIV}, and RC_{ID}.

Similar results were obtained when the statistical methods were compared in terms of their receiver operating characteristics (see Table 13). All statistical methods yielded relatively high AUC values (all p's < .001), which ranged from 0.97 to 0.88. The RCI_{HSU} had the greatest area under the ROC curve at 0.973 indicating that overall, this change score method was the most accurate relative to the diagnostic standard. The difference, however, between the RCI_{HSU} and other methods, such as $RCI_{CHEL}a$, the 3 SRB methods, and RCI_{SPEER} were small in magnitude. The methods with the smallest AUC values were the RC_{ID} and RC_{INDIV}. Overall, the findings from the correlational and AUC analyses regarding the sum of reliably changed scores did not lend support to the study hypothesis.

Table 14 contains data pertaining to the sensitivity and specificity of the change score methods using the entire CSHA neuropsychological test battery (a total of 16 measures/indices). Three different cut-off values are presented. The first set of cut-off values indicate the number of reliably deteriorated test scores in this sample that resulted in sensitivity rates at or greater than 95% along with the associated levels of specificity. The second set of cut-off values indicates the number of reliably deteriorated tests that resulted in 95% specificity in this sample along with the corresponding levels of sensitivity. Finally, the optimal cut-off values presented on the right side of Table 14 refer to those values that resulted in the highest specificity when sensitivity was at least 90%. As shown, sensitivity greater than or equal to 90% and specificity greater than or equal to 95% was obtained in this sample using a cut-off of 4 or more tests with the

RCI_{HSU}, 5 or more tests with the SRB_{MCS} and SRB_{CH}, and 6 or more tests with the RCI_{SPEER}. The RCI_{HSU}, which was identified as the best method in the correlational and AUC analyses, resulted in 90% sensitivity and 95% specificity rates using a cut-off of 4 or more tests; and 95% sensitivity with 90% specificity rates using a cut-off of 3 or more tests.

Hypothesis 13. It was expected that subjective reports of loss in cognition would not be strongly related to reliable deterioration in test scores. This question was examined using Phase E data from 228 participants with no reported cognitive decline and 25 individuals who reported decline. The findings revealed that no particular measure or statistical method was consistently associated with a subjective report of cognitive decline (all p's > 0.01). As expected, chi-square analyses showed no significant association between self-reported loss and test score change on the WMS Information subtest, BCR Acquisition, RAVLT Trial 1 and Delay, BVRT, WAIS-R Similarities, WAIS-R Comprehension, Token Test, verbal fluency (FAS and Animals), WAIS-R Block Design, and WAIS-R Digit Symbol subtests (all p's > 0.01). These measures were also associated with relatively low ORs (ORs < 3). In contrast, self-reported decline was significantly associated (p's < 0.01) with changes in BCR Trial 1, Retrieval, and Delay recall and changes in RAVLT Total recall using some, but not all, of the change score methods.

Overall, the BCR Retrieval score had the strongest relation with subjective cognitive loss of any measure (classification accuracy ranged from 67% to 77%). The methods with the highest ORs were RCI_{SPEER} (OR = 5.5 [95% CI = 2.2, 13.3]) followed by the SD, RCI, and $RCI_{JT}b$ methods (OR = 3.7 [95% CI = 1.5, 8.7]). The next highest

association was with the RCI_{HSU}, RC_{INDIV}, RC_{ID}, $RCI_{JT}a$, and $RCI_{CHEL}b$ methods. The remaining statistical methods had relatively weaker associations with ORs ranging from 2.0 to 2.3.

Test score change on BCR Trial 1 showed the greatest relation to self-reported change using the RCI and SD methods (OR = 3.3 [95% CI = 1.4, 7.8]) followed by RCI_{HSU} and RCI_{SPEER} (OR = 3.1 [95% CI = 1.3, 7.2]). The remaining methods showed weak or no association between subjectively reported loss and deteriorated change scores. Odds ratios could not be calculated for the RC_{ID} and RC_{INDIV}.

The RCI_{SPEER}, SRB_{CH}, and SRB_{MCS} showed a significant association between subjective report of decline and change in test performance on BCR Delayed recall (OR = 3.9 [95% CI = 1.7, 9.2]), followed by RCI_{HSU} and SRB_{MULT} (OR = 3.6 [95% CI = 1.5, 8.4]). Reliable change as calculated using the remaining methods had weak associations with reported loss on this memory index (ORs < 3.0). Odds ratios could not be calculated for the RC_{ID} and RC_{INDIV}.

The RAVLT Total score showed a significant association with loss using the SRB_{CH}, SRB_{MCS}, RCI_{HSU}, and RC_{INDIV} methods (OR = 3.9 [95% CI = 1.4, 10.5]), followed by RCI_{SPEER}, $RCI_{CHEL}b$, SD, and RCI (ORs ranging from 3.1 to 3.6). The remaining change score methods did not show a significant relation to self-reported loss using this measure.

In addition to the above evaluation of individual test/method combinations, the relation between cognitive loss noted by the participant and the sum of reliable changes across all 16 neuropsychological measures for each statistical method was also examined (see Table 15). Contrary to expectation, all methods but the RC_{ID} showed a statistically

significant association between subjectively-reported cognitive decline and the sum of reliably deteriorated scores (all p's < 0.05). The correlations were comparable across change score methods and were relatively weak in magnitude. They ranged from -0.189 to -0.115. The strength of the association was greatest for RC_{INDIV} and decreased in order of RCI_{SPEER}, SRB_{CH}, SRB_{MCS}, RCI_{HSU}, SRB_{MULT}, $RCI_{JT}b$, SD, RCI, $RCI_{JT}a$, $RCI_{JT}c$, $RCI_{CHEL}a$, $RCI_{CHEL}c$, $RCI_{CHEL}b$, and RC_{ID}.

An examination of the receiver operating characteristics (see Table 15) shows that all of the change score methods yielded AUC values in the 0.6 to 0.7 range. This range is usually associated with relatively weak or poor accuracy rates (recall that an AUC of 0.5, in this context, indicates no relation between test score change and self-reported change). The change score method with the greatest AUC value was the RC_{INDIV} at 0.667. The methods with the smallest, and also statistically non-significant, AUC values were the RC_{ID} and RCI_{CHEL} (using formulae b and c).

Overall, the above findings provide at least partial support to the study hypothesis. That is, the relation between self-reported loss and test score deterioration was not strong. As expected, change on most neuropsychological measures was not associated with self-reported decline. Change on some memory indices, however, was significantly related to self-reported loss using some change score methods though the magnitude of this association was relatively weak. Similar results were obtained using the sum of reliably changed test scores. As a pattern, the RC_{INDIV}, RCI_{SPEER}, RCI_{HSU}, and SRB methods yielded the strongest relation to self-reported change.

Hypothesis 14. It was expected that informants' reports of loss in cognition would be strongly related to reliable deterioration in test scores. Using Phase F data, this

phase involved an examination of data from 73 persons with informant-reported cognitive loss and 178 persons with no loss noted by an informant. Chi-square analyses revealed that no change score method was consistently associated with informant report of cognitive decline (all p's > 0.01). With regard to specific measures, the WMS Information subtest and the 4 BCR memory indices were consistently associated with informant ratings of decline in mental abilities across all statistical methods (all p's < 0.01). The RAVLT Total, WAIS-R Similarities, phonemic fluency (FAS), semantic fluency (Animals), and WAIS-R Digit Symbol also showed associations with informant ratings but only when using particular change score methods. Changes scores on the RAVLT Trial 1 and Delay recall, BVRT, WAIS-R Comprehension, Token Test, and WAIS-R Block Design showed no significant association with informants' reports of loss using any of the change score methods (all p's > 0.01). More detailed information follows for those measures found to have at least some relation to informants' ratings.

The WMS Information subtest resulted in classification accuracy that ranged from 60% to 72%. There were few differences among the methods in terms of their odds ratios, which ranged from 2 to 4. The RCI_{ID} yielded the highest odds ratio (OR = 4.0 [95% CI = 1.4, 11.8] but was associated with poor sensitivity (12%). SRB_{MCS} and SRB_{CH} were next best (OR = 3.5 [95% CI = 1.8, 6.7]) followed by the RCI_{HSU} (OR = 3.4 [95% CI = 1.8, 6.1]) and then the SRB_{MULT}, $RCI_{JT}b$, $RCI_{CHEL}b$, and $RCI_{CHEL}c$ (OR = 3.3 [95% CI = 1.7, 6.3]). The SD method demonstrated the next highest association followed by RCI_{SPEER}. The remaining statistical methods yielded ORs of 2.4 [95% CI = 1.3, 4.3].

Across all measures, BCR Trial 1 recall was most strongly associated with informants' ratings of loss (ORs ranging from 5 to 9). This relation was strongest using

the RCI_{HSU} (OR = 9.1 [95% CI = 4.5, 18.2]) followed by SRB_{CH}, SRB_{MCS}, RCI_{CHEL}a, b, and c (OR = 7.4 [95% CI = 3.3, 16.6]) and then SRB_{MULT} and RCI_{JT}a, b, and c methods (OR = 6.7 [95% CI = 3.5, 12.9]). The RCI_{SPEER} resulted in a stronger association between reliable deterioration and informant-reported loss than the RCI and SD methods (OR = 4.9 [95% CI = 2.7, 8.8]). Odds ratios could not be calculated for RC_{ID} and RC_{INDIV}.

BCR Retrieval change scores were related to informant reported change using several methods. Those methods resulting in the highest ORs were SRB_{CH}, SRB_{MCS}, and RCI_{CHEL}a and c (OR = 6.3 [95% CI = 3.2, 12.3]). These were followed closely by RCI_{SPEER}, RCI_{HSU}, and SRB_{MULT} (ORs > 5). The RCI_{JT}c yielded the next strongest association followed by RCI_{JT}a, RCI_{CHEL}b, RC_{ID} and RC_{INDIV} (ORs > 4). RCI_{JT}b, RCI, and the SD method showed the weakest associations (OR = 3.8 [95% CI = 2.1, 6.9]) between test score change and informants' ratings using this memory index.

Like the Retrieval score, BCR Acquisition change scores were also consistently related to informant reports of cognitive decline. The largest ORs were found using the RCI_{JT}b and RCI_{CHEL}b and c methods (OR = 8.5 [95% CI = 3.1, 22.7]). The RCI, RCI_{JT}a and c, RCI_{CHEL}a, and RC_{ID} yielded the next largest association (OR = 5.2 [95% CI = 2.3, 11.8]) followed by the RC_{INDIV} and SD methods. RCI_{HSU} and RCI_{SPEER} methods (OR = 4.4 [95% CI = 2.0, 9.5]) had a slightly greater association than the SRB methods, which on this measure, had the weakest relation with informant reported change.

On the BCR Delayed Recall, classification accuracy ranged from 70% to 74%. The SRB_{MULT}, RCI_{HSU}, SRB_{CH}, SRB_{MCS}, and RCI_{SPEER} were most accurate (ORs ranged from 5.3 to 5.7) and followed closely by the RCI_{JT} and RCI_{CHEL} methods (OR = 5.1 [95%

CI = 2.6, 9.9]). Change scores based on the SD and RCI methods had the weakest association (OR = 3.9 [95% CI = 2.2, 7.0]) with informant reported change on this measure. Odds ratios could not be calculated for RC_{ID} and RC_{INDIV}.

Across statistical methods, change scores on the RAVLT Trial 1 and Delayed Recall conditions were not significantly related to reports of cognitive loss (all p's > 0.01). However, the RAVLT Total score did show a significant association (p's < 0.01) with memory loss using the SRB_{CH}, SRB_{MCS}, and RCI_{HSU} methods (OR = 6.4 [95% CI = 2.7, 15.3]) followed by the SRB_{MULT} (OR = 5.3 [95% CI = 2.2, 12.4]). RAVLT Total recall change scores, as calculated using the remaining statistical methods, were not significantly associated with informant ratings of change.

Change on the WAIS-R Similarities subtest was generally not associated with informants' ratings of loss (all p's < 0.01 and ORs < 2). The only exception in this study was found using the RC_{ID}, which yielded an OR of 6.1 [95% CI = 1.5, 24.3]. Though the odds ratio was statistically significant, the RC_{ID} was associated with a particularly low sensitivity rate (9.6%) on this measure.

On a measure of phonemic fluency (FAS), classification accuracy ranged from 64% to 71%. Odds ratios were generally in the 2 to 3 range. Change scores calculated using the RCI_{HSU} method resulted in the strongest association with cognitive loss reported by informants (OR = 3.8 [95% CI = 1.4, 10.5]), followed by the SD method, RCI_{SPEER}. There were few differences among the remaining methods (ORs < 3).

Similar results were obtained using change scores on the semantic fluency task. Change scores calculated using the SRB_{CH}, SRB_{MCS}, and RCI_{HSU} (OR = 5.3 [95% CI = 2.0, 14.1]) were most strongly associated with informant reports of decline, followed

closely by the SRB_{MULT} method (OR = 4.6 [95% CI = 1.6, 13.2]). The remaining statistical methods yielded non-significant associations with ORs ranging from 2.1 to 2.7.

WAIS-R Digit Symbol change scores were related to informant rating of change using a few statistical methods (ORs ranging from 2 to 7). The SRB_{MULT} (OR = 7.0 [95% CI = 2.3, 21.0]) and RCI_{HSU} (OR = 6.3 [95% CI = 2.5, 15.9]) were the most strongly associated methods followed by SD (OR = 6.2 [95% CI = 2.8, 13.7]), $RCI_{CHEL}a$ (OR = 5.5 [95% CI = 2.0, 14.7]), and the remaining regression-based methods (OR = 4.4 [95% CI = 1.6, 12.1]). The next best methods included the $RCI_{JT}a$ and c, RCI_{SPEER}, RCI, and $RCI_{CHEL}b$. Changes scores based on the $RCI_{CHEL}c$ and $RCI_{JT}b$ were not significantly associated with informants' ratings of change.

The analyses in Phase F also included an examination of informants' ratings of change and the sum of reliable changes on neuropsychological measures for each statistical method. These findings are summarized in Table 16. Informants' reports of loss were significantly associated with reliable deterioration across all statistical methods when the number of reliably declined test scores were summed (all p's < 0.001). The correlations, however, were generally low and comparable among change score methods with values in the 0.2 to 0.3 range. The SRB_{CH} yielded the strongest correlation (Spearman's rho = -0.346) followed by the SRB_{MCS}, and RCI_{HSU} methods. The remaining correlations, in decreasing order, were SRB_{MULT}, SD, RCI_{SPEER}, $RCI_{CHEL}a$, $RCI_{JT}a$, RCI, $RCI_{JT}c$, $RCI_{CHEL}c$, $RCI_{JT}b$, $RCI_{CHEL}b$, RC_{INDIV}, and RC_{ID}, which yielded the weakest correlation (Spearman's rho = -0.229).

Similar results were obtained when the statistical methods were compared in terms of their receiver operating characteristics (see Table 16). All of the change score

methods yielded statistically significant AUC values in the 0.6 to 0.7 range. This range is usually associated with low to fair accuracy rates. The SRB_{CH} had the greatest area under the ROC curve at 0.713. The method with the lowest AUC value was the RC_{ID} at 0.640.

These data, on the whole, provided minimal support for the study hypothesis as the relation between change and informants' ratings was statistically significant but not strong. The relation was greatest using certain memory indices (e.g., WMS Information and BCR) and some other measures (e.g., verbal fluency and psychomotor speed) depending on the change score method employed to classify change.

Hypothesis 15. According to this hypothesis, clinicians' ratings of loss were expected to be strongly associated with reliable deterioration in test scores. In Phase G, clinicians' ratings data from 51 individuals deemed to have cognitive loss were compared to 173 persons deemed to have no loss. Chi-square analyses revealed that no change score method was consistently associated with clinician report of cognitive decline (all p's > 0.01). With regard to specific measures, the WMS Information subtest and the 4 BCR memory indices were consistently associated with clinicians' ratings of cognitive decline across all statistical methods (all p's < 0.01). The RAVLT (Trial 1 and Total), BVRT, Similarities, and semantic fluency also showed associations with clinicians' ratings but only with particular statistical methods. The RAVLT Delay recall, WAIS-R Comprehension, Token Test, phonemic fluency, WAIS-R Block Design, and WAIS-R Digit Symbol showed no significant association with ratings of loss (all p's > 0.01) regardless of the change score method that was employed. More detailed information follows for those measures found to have at least some relation to clinicians' ratings.

Test score change on the WMS Information subtest resulted in classification accuracy that ranged from 61% to 79%. Most ORs were in the 4 to 6 range. The RCI_{ID} yielded the highest odds ratio at 22.9 [95% CI = 2.7, 195.4] but was associated with poor sensitivity (12%). The $RCI_{JT}b$, $RCI_{CHEL}b$, and $RCI_{CHEL}c$ had the next highest odds ratios (OR = 6.8 [95% CI = 2.3, 19.8]) followed by SRB_{MCS} and SRB_{CH} (OR = 5.6 [95% CI = 2.4, 13.0]) and RCI_{HSU} (OR = 4.9 [95% CI = 2.3, 10.2]). All but one of the remaining statistical methods yielded odds ratios ranging from 4.2 to 4.4. The SD method resulted in the lowest association with clinicians' ratings of loss (OR = 2.7 [95% CI = 1.4, 5.1]).

Trial 1 recall on the BCR was most strongly associated with clinicians' ratings of loss using the $RCI_{CHEL}a$, b, and c (OR = 7.4 [95% CI = 2.6, 21.2]) and SRB_{MULT} (OR = 7.0 [95% CI = 2.6, 19.0]). The SRB_{CH}, SRB_{MCS} were next best (OR = 5.9 [95% CI = 2.4, 14.4]) followed by the $RCI_{JT}a$, b, and c methods (OR = 5.6 [95% CI = 2.6, 12.1]). The RCI_{HSU} (OR = 4.9 [95% CI = 2.2, 11.3]) resulted in a stronger association between reliable deterioration and clinicians' ratings than the RCI and SD methods (OR = 3.4 [95% CI = 1.7, 6.5]), which in turned fared slightly better than RCI_{SPEER} (OR = 3.1 [95% CI = 1.5, 6.5]). Odds ratios could not be calculated for RC_{ID} and RC_{INDIV}.

Among all measures, BCR Retrieval change scores were more strongly related to clinician-reported change with classification accuracy ranging from 76% to 80%. Those methods resulting in the highest odds ratio were SRB_{CH} and SRB_{MCS} (OR = 9.6 [95% CI = 4.0, 23.2]). This was followed closely by $RCI_{CHEL}a$ and c (OR = 8.6 [95% CI = 3.6, 20.2]), SRB_{MULT} (OR = 8.4 [95% CI = 3.7, 19.4]), and $RCI_{JT}c$ and RCI_{HSU} (OR = 8.3 [95% CI = 3.8, 18.1]). The next strongest associations were $RCI_{JT}a$, $RCI_{CHEL}b$, and RC_{ID} (OR = 7.4 [95% CI = 3.5, 15.4]). The $RCI_{JT}b$, RCI_{SPEER} RCI, and the SD method fared

slightly better (OR = 6.8 [95% CI = 3.5, 13.5]) than the RC_{INDIV} (OR = 6.1 [95% CI = 3.1, 12.2]) using this memory measure.

Like the Retrieval score, BCR Acquisition change scores were also consistently related to clinicians' ratings of cognitive loss though the sensitivity was low (13% to 35%). The largest odds ratios with this measure were found using the RCI_{JT}b and RCI_{CHEL}b and c methods (OR = 13.1 [95% CI = 2.6, 65.4]). The RCI_{HSU} and RCI_{SPEER} methods resulted in high odds ratios of 11.2 [95% CI = 3.4, 37.0] and were followed by RCI, RCI_{JT}a and c, RCI_{CHEL}a, and RC_{ID} (OR = 9.9 [95% CI = 3.0, 33.3]). Relatively weaker associations were found using the SRB methods (OR = 4.2 [95% CI = 2.0, 9.0]). The RC_{INDIV} and SD methods (OR = 3.9 [95% CI = 1.8, 8.7] had the weakest relation with clinician-reported change using the BCR Acquisition test-retest scores.

On the BCR Delayed Recall, classification accuracy ranged from 74% to 81%. ORs ranged from 4 to 9. The RCI_{HSU} produced changes scores on this measure that were most strongly associated to clinicians' ratings (OR = 9.7 [95% CI = 4.5, 21.1]). The SRB_{MULT} was next best (OR = 8.6 [95% CI 4.1, 18.0]) followed by all of the RCI_{JT} and RCI_{CHEL} methods (OR = 8.3 [95% CI = 3.7, 18.8]). The SRB_{CH}, SRB_{MCS}, and RCI_{SPEER} yielded stronger associations (OR = 7.6 [95% CI = 3.7, 15.7]) than change scores based on the SD and RCI methods (OR = 4.7 [95% CI = 2.4, 9.3]).

As previously mentioned, change scores on the remaining neuropsychological measures showed associations with clinicians' ratings of loss using only particular statistical methods. Only the SRB_{CH} and SRB_{MCS} methods (OR = 3.7 [95% CI = 1.3, 10.4]) showed a statistically significant relation between test score change on the RAVLT Trial 1 and the ratings of CSHA clinicians (all other p's > 0.01). Using the total

score from the RAVLT, a significant association with ratings of loss was found with the SRB_{MULT} (OR = 14.7 [95% CI = 4.6, 47.4]), the SRB_{CH}, SRB_{MCS}, and RCI_{HSU} methods (OR = 9.5 [95% CI = 3.1, 28.6]) and the RCI_{CHEL}a and c methods (OR = 7.8 [95% CI = 2.8, 22.3]). Weaker associations emerged using the RCI_{SPEER} (OR = 5.3 [95% CI = 2.5, 11.6]), RCI_{JT}a and c (OR = 4.2 [95% CI = 1.8, 9.8]), and RCI and SD methods (OR = 2.8 [95% CI = 1.4, 5.6]). Ratings of loss and change on the BVRT were related using the SRB_{CH} and SRB_{MCS} (OR = 5.3 [95% CI = 2.3, 12.3]), SRB_{MULT} (OR = 4.7 [95% CI = 2.0, 11.3]), and RCI_{HSU} and RCI_{SPEER} methods (OR = 4.3 [95% CI = 1.8, 10.3]). Using the WAIS-R Similarities subtest, the RC_{ID} (OR = 11.4 [95% CI = 2.2, 58.4]) showed strong associations between change on test scores and clinicians' ratings although the sensitivity was notably poor (11%). Associations using the RC_{INDIV}, RCI_{CHEL}a and c, RCI_{JT}a and c, and the SD method were relatively weak (OR = 3.6 [95% CI = 1.4, 9.0]). Finally, change on the semantic fluency task was related to clinicians' ratings of loss using the SRB_{CH}, SRB_{MCS}, and RCI_{HSU} (OR = 5.5 [95% CI = 1.5, 20.3]) methods followed by the RCI_{SPEER} (OR = 4.1 [95% CI = 1.8, 9.6]).

The analyses in Phase G included an examination of clinicians' ratings of change and the sum of reliable changes on the 16 neuropsychological measures for each statistical method (see Table 17). All change score methods were significantly related to the sum of reliable changes (all p's < 0.001). The strength of the correlations was comparable across methods and generally ranged from 0.3 to 0.4. The SRB_{CH} yielded the strongest correlation (Spearman's rho = -0.431) followed by the SRB_{MCS}, RCI_{SPEER}, SRB_{MULT}, and RCI_{HSU} methods. The remaining correlations, in decreasing order, were RCI_{JT}c, RCI_{JT}a, RCI_{JT}b, SD, RCI_{CHEL}a, RC_{INDIV}, RCI, RCI_{CHEL}c, RCI_{CHEL}b, and RC_{ID}.

Inspection of Table 17 also shows that the AUC values from Phase G data were all statistically significant (all p's < 0.001). Most of the AUC values were in the 0.7 to 0.8 range, indicating that reliably deteriorated test scores could distinguish, to a fair degree, between clinician's ratings of loss versus no loss. RCI_{SPEER} had the greatest area under the ROC curve at 0.788 followed by SRB_{CH} and SRB_{MCS}. The method with the least area under the ROC curve was the RC_{ID} at 0.680.

Overall, the findings from Phase G provided moderate support for the study hypothesis regarding the relation between clinicians' ratings and change score classification. The relation was greatest using certain memory indices (e.g., WMS Information and BCR). Other memory indices (e.g., RAVLT and BVRT) along with WAIS-R Similarities and semantic fluency only demonstrated a relation between test score change and clinicians' ratings using specific methods. Taken together, test score change across the entire neuropsychological test battery was significantly related to clinicians' ratings of loss.

Discussion

This study focused on neuropsychological change in older adults and addressed two main questions: 1) how much change is normal using various change score methods over varying lengths of time in cognitively intact participants, and 2) which methods and measures are useful in determining reliable change in performance that corresponds to indices of clinically significant change? These questions were considered using two different samples of older adults from the CSHA using multiple neuropsychological measures, 15 different change score methods (including various permutations of

formulae), and 4 different indices of clinically significant change. The findings provided mixed support for the study hypotheses. What follows is a brief summary of the main findings and then a more detailed section addressing some of the study hypotheses.

With regard to the first research question, the results indicated that there was a substantial degree of variation in cognitive test performance over time among older adults. Though performance varied considerably, the data supported the notion that "normal" change can be accurately classified using change score methods. The RCI_{HSU}, RCI_{CHEL} (formulae a and c), all 3 SRB methods, and to a lesser extent the RCI_{JT} (formulae a and c), were accurate in defining normal change among the cognitively healthy older adult samples. In contrast, the $RCI_{JT/CHEL}$ (formula b), RCI_{SPEER}, RCI, SD, RC_{ID}, and RC_{INDIV} were consistently less accurate in identifying normal change. When patterns of decline consistent with dementia were examined (as compared to change on individual tests), all of the change score methods yielded acceptable accuracy rates over the shorter interval of a few months but the RCI_{CHEL} (formulae a and c), RCI_{HSU}, 3 SRB methods stood out as being the most accurate over a 5-year span. With the exception of RC_{ID} and RC_{INDIV}, all of the statistical methods adequately captured "abnormal" change in persons who developed dementia after several years.

The answer to the second research question regarding the utility of change score methods and neuropsychological measures depended on how clinically significant change was operationalized. Using a "gold standard" of consensus diagnostic opinion, change scores computed using the RCI_{HSU} and 3 SRB methods were consistently accurate and most strongly related to change in diagnostic status (i.e., a transition from NCI at baseline to a diagnosis of dementia at follow-up). The most discriminating neuropsychological

measure, overall, was the BCR (particularly the total free recall score) though other tests including the WMS Information subtest, the BVRT, WAIS-R Comprehension, the Token Test, and a measure of phonemic fluency (FAS) were also useful in this regard. When the entire battery was examined and the sum of measures that demonstrated reliable test score deterioration was calculated, diagnostic change was significantly associated with all of the change score methods but most strongly associated with the RCI_{HSU}. Reliable declines on test scores were moderately related to clinicians' ratings of change, mildly related to informant's ratings of change, and weakly related to subjective ratings of change in this study. Memory test score changes were most strongly associated with cognitive loss defined using perspectives other than diagnostic change. Other non-memory measures were also related to judgments of loss by informants and clinicians but only when using specific change score methods.

Characteristics of older adults' performance. Normal variation in test performance and in test score differences were observed among all CSHA measures but the BCR Acquisition score. Scores on this latter measure were non-normally distributed, evidenced a ceiling effect, and showed relatively little variation in test performance over time suggesting that this index in particular would be poor in change score measurement. The remaining measures, in contrast, demonstrated at least fair amounts of intraindividual change and/or group-level change.

In this study, the test-retest reliabilities of the neuropsychological measures ranged considerably but were comparable to those derived from other studies using older adult samples (Paolo & Ryan, 1993; Spreen & Strauss, 1998). Single trial indices on the BCR and RAVLT memory measures (e.g., Trial 1 recall and Delayed recall) had

relatively low correlations but aggregate memory indices (e.g., BCR Retrieval and RAVLT Total) had more desirable test-retest reliability coefficients, as one might expect. The remaining measures in the CSHA neuropsychological test battery evidenced moderate to high correlations over a brief interval spanning weeks to months as well as over a longer interval of 5 years. The strongest test-retest correlations were obtained for the speeded verbal fluency and psychomotor tasks.

Varying the length of the retest interval (as well as some of the sample characteristics like age and sample size) had some effect on test-retest reliability. With the exception of the WAIS-R Similarities subtest, test-retest correlations tended to be slightly higher for most neuropsychological measures over the shorter follow-up period. The length of the retest interval, in contrast, had little effect on the coefficients for the BCR indices and the WAIS-R Digit Symbol subtest, indicating that changes in rank-level performance were not a function of this variable.

Over a 5-year interval, the reliability of the difference (r_{DD}) scores that were calculated for RC_{ID} and RC_{INDIV} were relatively high for the BCR Acquisition, RAVLT Total and Delay indices, and the Token Test. The WAIS-R Similarities and BCR Retrieval scores were associated with moderate r_{DD}s and the remaining measures had disappointingly low r_{DD}s. Low r_{DD}s (i.e., r_{DD}s ≤ 0.40) were found for virtually all measures over the shorter interval. As previously noted, this had a large impact on the study by limiting the ability to calculate change scores using the RC_{INDIV}. The reason why many r_{DD}s were so low is not clear. The fact that lower r_{DD}s were obtained for most measures in Phase A compared to Phase B suggests that: 1) either larger samples (i.e., $n > 30$) are needed to obtain reasonable estimates of r_{DD}, or 2) test-retest intervals greater

than a few months are needed to obtain adequate levels of variability in test score differences. One can also speculate that low r_{DD} values may have been low due to a restricted range of variability inherent in the measures, for example, due to floor/ceiling effects and the use of short forms (i.e., some of the WAIS-R measures). It is also possible that age may have been a factor although the substantial variability among older adults' performance makes it less likely that age limited the reliability of test score differences. The manner in which r_{DD} was defined using various coefficients (e.g., Guttman reliability coefficients from time 1 and time 2) that were derived from relatively small samples may have had a greater impact on the adequacy of the r_{DD} values in this study.

Group-level change findings. Older adults' test performance, on average, remained stable over an interval of a few months but significantly declined over 5 years on specific neuropsychological measures. As predicted (Hypothesis 2), NCI participants who were followed over the longer interval showed either stable or deteriorated mean test performance. Significant declines were found on several memory tests as well as on timed measures that required quick responses. These findings are consistent with both cross-sectional and longitudinal studies of older adults that have demonstrated age-related changes in memory and psychomotor speed (Albert, 1994; Colsher & Wallace, 1991; Schaie, 1989). In contrast, performance on language, abstraction, and judgment measures did not substantially drop in this sample over the same span.

Practice effects, as defined by significant group-level improvements in test performance, were not the norm among adults over age 65 in this study. Over the brief interval of a few months, the expected practice effects were not found (Hypothesis 1).

Subtle practice effects may have been present, as indicated by mean improvement in most follow-up test scores, but the extent of these changes was not pronounced. The study design in Phase A, including the small sample size and the length of the "brief" test-retest interval, in addition to the conservative alpha level used in the statistical analyses may have made it more difficult to detect practice effects among older adults. A shorter interval on the scale of days or a couple of weeks may have generated different results. The lack of practice effects in Phase A might also be explained by the age of the participants, which was 75 and greater. Previous research has shown that the effects of practice diminish quickly in older adults around this age (Mitrushina & Satz, 1991; Ryan et al., 1992). The data from this study, however, do not support this finding. All individuals from Phase A (who ranged in age from 75 to 97 years) showed a mixture of improvements and declines and there was no association between age and change in test score performance demonstrated using any measure.

Over the longer 5-year interval, practice effects (i.e., mean improvements) were absent and the pattern of mean decline that was observed on some measures was likely due to mild age-related deterioration in psychomotor speed and memory retrieval. These group-level analyses, however, obscure individual differences. Higher scores after 5-year follow-up were shown by some participants in the CSHA on each and every one of the neuropsychological measures. If this improvement can be attributed to practice effects (rather than factors such as regression to the mean), it suggests that the effects of repeated exposure to a test, or to testing in general, may be specific to the individual and not simply a function of age.

The findings from this study provided at least partial support for Hypothesis 3, which stated that persons who developed dementia by follow-up would demonstrate significant global deterioration across all neuropsychological measures relative to those who remained cognitively healthy. Significant declines were evident on many measures and non-significant trends in the expected direction were found on the remaining tests. These results are not surprising since dementia, by definition, involves deterioration in multiple areas of cognition. The findings, however, indicate that some measures are more sensitive than others to the changes that occur as part of a progressive dementia. Important factors to consider are the psychometric properties of the individual test and whether or not floor effects are present. It is also important to note that not all cognitive abilities decline early, or at the same rate, in persons with dementia. Declines in memory, verbal fluency, and abstraction and reasoning abilities, as shown in this study, are often among the first changes in persons with dementia, particularly Alzheimer Disease (Albert, Moss, Tanzi, & Jones, 2002; Bondi et al., 1994; Jacobs et al., 1995; Masur et al., 1994).

Individual-level change among cognitively healthy older adults. Normal change in test performance, the focus of Phases A and B, was expected to be best classified using the SRB methods, followed by the RC_{INDIV}, RC_{ID}, RCI_{CHEL}, RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and finally, the SD method (Hypotheses 4 and 7). The data from the short (Phase A) and long (Phase B) retest intervals did not support these hypotheses. In Phase A, the RCI_{HSU} was the most accurate method followed closely by the RCI_{JT} and RCI_{CHEL}. In Phase B, the most accurate method was the RCI_{CHEL} and the RCI_{JT} and RCI_{HSU} showed a moderately-high level of accuracy. As a pattern, these RCI methods performed better, or

at least as well, as the 3 SRB methods in classifying normal change. Other RCI methods, such as the RC_{ID}, RC_{INDIV}, and RCI_{SPEER} performed worse than expected and often yielded lower accuracy rates than the SD method.

The superiority of the RCI_{CHEL} (using formulae a and c) over more sophisticated SRB methods in classifying normal change, as found in this study, has also been observed in investigations by Temkin, Heaton, and their colleagues (Heaton et al., 2001; Temkin et al., 1999). The RCI_{JT} (using formulae a and c), though accurate in Phase A, performed slightly less well than the SRB methods in Phase B. This pattern is also consistent with the results of previous research (Temkin et al., 1999). Recall that the RCI_{CHEL} is a modification of the RCI_{JT} method that accounts for practice (or aging) effects by adding (or subtracting) a constant. The constant reflects the mean test score change in the sample. In Phase A, the RCI_{CHEL} and RCI_{JT} fared comparably, as one might expect, since there were no significant practice effects in neuropsychological test performance and no aging effects were expected given the brief interval. With a longer retest interval where practice effects were absent and aging effects were evident on some measures, the RCI_{CHEL} performed better than the RCI_{JT}. This finding highlights the importance of correcting for age effects with older adults, at least on some measures.

The findings from this study demonstrated that the formulae employed in calculating the SED for the RCI_{JT} and RCI_{CHEL} significantly influenced the utility of these change score methods. As expected, classification accuracy was consistently and substantially higher using formula c (i.e., where the SED was directed observed from the longitudinal data) relative to formula b. An unexpected finding was that formula a (i.e., which estimated the SED by multiplying the SEM at time 1 by 1.414) was also better

overall than formula b and generally produced results that were consistent or just slightly worse than those obtained using formula c. The use of formula b, particularly with RCI_{JT}, was poorer than most other change score methods due to its tendency to classify too many cognitively healthy older adults as showing reliable deterioration. It is unclear why the use of formula b, a relatively sophisticated method of estimating the SED which takes into account both the SEM at time 1 and a different SEM at time 2, consistently rendered lower accuracy rates over both test-retest intervals in classifying normal change in older adults. This issue needs to be investigated in other studies. If these findings are replicated, they may have implications for studies that have employed formula b in calculating the SED as advocated by Iverson (Iverson, 1999, 2000; Iverson & Green, 2001).

An unexpected finding in this study was the high accuracy rate associated with the RCI_{HSU}. It was the most accurate method for classifying normal change among cognitively healthy older adults over a short interval and retained moderate-high accuracy over the 5-year period. Recall that the RCI_{HSU} employs a residualized gain score to correct for the effects of regression to the mean. The effectiveness of the RCI_{HSU} in this study suggests that this factor may be important when interpreting older adult serial assessment data, particularly in the absence of practice effects. Compared to other change score methods, the RCI_{HSU} has received little attention in previous research. Only one study has previously examined it, and in that study Speer (1995) showed that the RCI_{HSU} produced discordant classifications relative to RCI_{SPEER} and RCI_{JT} and was more likely to classify persons as deteriorated on a scale of well-being. The conclusions drawn by Speer regarding the utility of the RCI_{HSU} are not consistent with the findings from the

current study which showed that this method was quite accurate in classifying normal change among cognitive healthy adults.

The three SRB methods, which were expected to perform the best, did not do so in Phase A. They were, however, quite accurate over the longer interval. One can speculate that this improvement in classification accuracy reflected the use of a larger sample of cognitively healthy older adults in Phase B relative to Phase A (166 versus 30 individuals), which in turn allowed the derivation of more accurate equations. If true, this finding suggests that the use of large samples may be paramount when generating regression-equations for change score purposes. Overall, the SRB methods produced similar results to each other. The simple regression methods (i.e., SRB_{MCS} and SRB_{CH}) yielded identical or nearly identical results on most measures over the short and long retest intervals. Compared to the SRB_{MCS}, recall that the SRB_{CH} is a more sophisticated and technically correct change score method that makes adjustments for using sample statistics to estimate population parameters. It could be argued that the benefit of using the SRB_{CH} might not have been appreciated in Phase B since the sample size was fairly large (n = 166) and may have provided a reasonable approximation of the population parameters. This argument, however, does not explain the similarities between SRB_{MCS} and SRB_{CH} in Phase A, which were based on data from only 30 participants. Change scores based on multiple regression equations (SRB_{MULT}) that accounted for age, gender, and education did not substantially improve classification accuracy over and above the use of baseline test performance in this study. Initial test performance, as shown by Temkin et al. (1999), by far accounted for the greatest proportion of variance in follow-up test performance. It is possible that the SRB_{MULT} may have performed better with

different predictors though the factors selected were those that traditionally account for the most variance in neuropsychological test performance.

Though purported to correct for several forms of errors and bias (Hageman & Arrindell, 1993, 1999b), the RC_{ID} and RC_{INDIV} did not fare well in classifying normal change in this study. The RC_{ID} was particularly conservative and often classified zero persons as changed over both the short and long retest intervals. Though it could be argued that this was technically accurate in a sample of persons with no known cognitive change, this method as shown in subsequent analyses was also poor in classifying clinically significant change. Both the RC_{ID} and RC_{INDIV} methods also proved to be complex to calculate given the number of required statistics including Guttman's reliability coefficients and the reliability of the difference score. Low r_{DD}s precluded the calculation of RC_{INDIV} for several measures, and as previously noted, there may have been a number of reasons for this finding. Calculating RC_{INDIV}, however, when r_{DD} was less than .40 (the cutoff specified by Hageman & Arrindell) also yielded poor accuracy rates. In this study, the RC_{INDIV} was reasonably accurate in classifying change using the BVRT and Token Test over the long interval but tended to overestimate deterioration on other measures. Although the RC_{INDIV} was not as thoroughly examined as other methods in this study, the findings suggest that the use of the phi-strategy does not improve accuracy over more traditional alpha-based change score methods.

As expected, the original RCI and RCI_{SPEER}, which employ the SEM as an error term, generally resulted in lower accuracy rates than RCI_{JT}, RCI_{CHEL}, and RCI_{HSU} in classifying normal change. The original RCI consistently outperformed the RCI_{SPEER}. The latter was prone to overestimate improvement over the short interval and misclassify

change in both directions over the longer retest interval. Typically, the RCI performed at the same level as the SD method.

The SD method was less accurate than several other change score methods but, contrary to the study hypothesis, consistently fared better than the RC_{ID} and RC_{INDIV}. It also fared better than the $RCI_{JT}b$ and RCI_{SPEER} over the long retest interval. The SD method classified too many individuals as improved regardless of the retest interval length, but was particularly prone to overclassify deterioration in test performance over the 5-year interval. This latter finding is consistent with previous research using measures with moderate test-retest reliability (Bruggemans et al., 1997; Kneebone et al., 1998).

Previous research (e.g., Temkin et al., 1999) has shown that the most accurate change score methods were those associated with the narrowest prediction intervals. Intuitively this is appealing since the less error associated with a measure, the narrower the interval that demarcates "normal" from "abnormal" change. The findings from this study, however, were not consistent with the expected results (Hypotheses 5 and 8) and did not support a relation between width of prediction interval and accuracy rate. In contrast to Temkin et al. (1999), smaller prediction intervals were not associated with higher classification accuracy and the SRB methods did not consistently result in the smallest prediction intervals. The most accurate methods in Phase A (RCI_{HSU}), Phase B ($RCI_{CHEL}c$), and Phase C ($RCI_{JT}b$) were associated with moderate to large prediction intervals as were the SRB methods. In contrast, the change score methods that consistently yielded the smallest prediction intervals (i.e., RC_{INDIV}, SD, RCI, and RCI_{SPEER}) were also the least inaccurate in classifying normal change. The difference

between the findings from this study and those of Temkin et al. (1999) may be due to differences in the neuropsychological measures used, the length of the test-retest interval, and the mean age of the samples studied. It is plausible that greater margins of error are needed to classify normal change in older adults given the considerable degree of variability in their performance (Ivnik et al., 2000).

As specified in Hypotheses 6 and 9, no statistical method was expected to show a pattern of reliable deterioration on measures of memory and one (or more) other cognitive domains in a sample with no known cognitive impairment at a rate greater than chance. The data from Phase A supported this hypothesis, but the data from Phase B did not. There are several implications from these results. First, these findings suggest that any of the change score methods can satisfactorily classify normal change in older adults over a brief period of time when reliable changes across multiple measures are considered. Second, only select methods maintain an acceptable level of accuracy in classifying normal change over longer test-retest intervals. The $RCI_{CHEL}a$, $RCI_{CHEL}c$, RC_{ID}, RC_{INDIV}, RC_{HSU}, and the 3 SRB methods were all associated with below chance misclassification rates in this study, but it is likely that the RC_{ID} and RC_{INDIV} methods were included in this group as an artifact of their conservative nature. The $RCI_{JT}b$, RCI, and SD methods, in contrast, each had a particularly high risk of misclassifying individuals in this sample as showing more deterioration than expected. A third implication of these findings relates to the importance of considering chance rates of misclassification. As the number of tests administered increases, the likelihood of finding evidence of reliable change also increases due to chance factors. Without lending consideration to familywise error rates, as noted by Keith et al. (2002), one might find

evidence of reliable deterioration on at least a few test scores without there being a true decrement in underlying cognitive ability. The data from this study illustrate this point. Using a battery of 16 measures, all of the statistical methods classified individuals who were NCI at both CSHA-1 and CSHA-2 as showing reliable decrements in multiple areas of cognition. Ivnik et al. (2000) reported similar results in their sample of older adults. Only by making reference to the chance misclassification rate did some of the statistical methods emerge as being useful in distinguishing between normal change and change consistent with a dementia pattern.

Individual-level change among older adults who progressed to dementia. It was hypothesized that all persons who progressed to dementia would evidence reliable deterioration on a measure(s) of memory plus one other cognitive domain(s). The results from Phase C partially supported this hypothesis. All of the statistical methods, with the exception of RC_{INDIV} and RC_{ID}, classified change in NCI persons who had progressed to a dementia with accuracy rates ranging from 80% to 100%. The RC_{INDIV} and RC_{ID}, which were too conservative or prone to errors in classifying normal change, were shown in Phase C to be less sensitive to dementia-related changes in test scores than the other methods. Taken together, these data suggest that RC_{INDIV} and RC_{ID} may be too conservative and/or inaccurate to be usefully applied in the serial neuropsychological assessment of older adults.

The $RCI_{JT}b$, $RCI_{CHEL}b$, RCI, RCI_{SPEER}, and SD methods, which were of suboptimal accuracy in Phases A and B, were the most accurate in Phase C. This finding suggests that these methods may simply be more sensitive to deterioration in test performance than other change score methods. Unfortunately, their tendency to

inaccurately classify persons with no significant cognitive decline (as shown in Phases A and B) may contraindicate their clinical use, particularly when the RCI_{HSU}, RCI_{CHEL}, and SRB methods performed quite well in classifying both normal change and "abnormal" change associated with progressive dementia.

Individual-level change and its relation to clinically-significant change. Overall, there was a fair amount of consistency and overlap among the statistical methods and their relation to clinically significant change. The change score methods, however, differed in their utility across the various phases of the study and were not equivalent in their accuracy rates. Using consensus diagnostic opinion as the "gold standard," there was a moderate and significant relation with test score changes. These findings are not consistent with the those of Ivnik et al. (2000) who used a similar standard and concluded that reliable change in test scores (using the RCI_{CHEL}) did not contribute to dementia diagnosis in older adults beyond chance levels. These discrepant findings may reflect differences in the samples that were examined, the design of the study, and the measures that were used. Recall that Ivnik et al.'s (2000) did not examine diagnostic change by using only persons who started out with no cognitive impairment at baseline testing and then forming groups based on whether or not they progressed to dementia at follow-up as was done in this study. Rather, they distinguished between normal and cognitively impaired groups at follow-up and determined that there were no differences in their change scores over time. Since it is not clear that the individuals in the two groups were cognitively healthy at baseline, it is impossible to determine the full extent of change associated with progression to dementia from their study. A related point is that Ivnik et al.'s (2000) cognitively impaired group was heterogeneous and included persons who did

not meet criteria for dementia making their results more difficult to interpret. Second, Ivnik et al. (2000) only examined 5 cognitive factors (i.e., MCFS) that are aggregate measures derived from the WAIS-R, WMS-R, and RAVLT. It is possible that these aggregate measures may have obscured decline on any one particular test thereby lowering the sensitivity of the change score approach. The findings from this study also indicate that the choice of memory measure may be a factor. In this study, the BCR indices showed much greater relation to clinically significant change than those from the RAVLT. Perhaps the use of different memory measures in the Mayo study may have produced different results. Finally, the present study and Ivnik et al.'s (2000) study differed with regard to the length of the test-retest interval. The CSHA involved examining change over a 5-year interval and may have been more likely to capture change than the Mayo study, which had a shorter test-retest interval of approximately 3 years.

Although change scores were generally found to be related to diagnostic change, there was no support for the expected pattern of superiority among change score methods using any of the CSHA neuropsychological measures. That is, no measure corresponded to the pattern of having the highest classification accuracy with the SRB_{CH} and SRB_{MULT} followed by SRB_{MCS}, RC_{INDIV}, RC_{ID}, RCI_{CHEL}, RCI_{HSU}, RCI_{JT}, RCI_{SPEER}, RCI, and the SD method (Hypothesis 11). This pattern also did not appear when the sum of reliably changed scores was examined (Hypothesis 12). Overall, the results from this study suggest that the RCI_{HSU} would be the best change score method to use when attempting to make an accurate diagnostic discrimination based on reliable change scores in older adults. The RCI_{CHEL} (using formula a or c) provides a close second choice. The SRB

methods were also highly accurate in this regard with the SRB_{CH} slightly outperforming the SRB_{MCS} and the latter faring slightly better than the SRB_{MULT}.

If the CSHA neuropsychological battery is used for the purpose of serial assessment, a clinician may use the data in Table 14 to determine the best cutoff (i.e., number of reliably changed scores) needed to increase the accuracy of a dementia diagnosis. The study findings also suggest that certain measures may be more important than others in detecting the changes that accompany a dementia process. The BCR stood out as the most discriminating measure in the CSHA test battery along with the WMS Information subtest. The combination of measures and statistical methods that yielded the best diagnostic accuracy (by consensus opinion) was the BCR Retrieval score using the RCI_{HSU} method and the BCR Delayed recall score using either the SRB_{CH} or SRB_{MCS} method. Both combinations resulted in an 89% overall correct classification rate with 89% sensitivity and 89% specificity. The odds ratio associated with the first combination was 68.2 [95% CI, 14.4 – 322.6] and the second combination was 64.0 [95% CI, 13.6 – 301.4]. This indicates that persons who showed reliable deterioration on the BCR Retrieval or Delay recall condition using these methods were over 60 times more likely (or at least 13 times more likely with 95% confidence) to receive a diagnosis of dementia at follow-up than remain NCI. In clinical practice, the use of the BCR Retrieval with the RCI_{HSU} may be the better option given the higher reliability of this index relative to Delayed recall. These findings are promising and suggest that a relatively easy to administer test of memory may be diagnostically useful in serial neuropsychological assessments with older adults. Using CSHA data for reference values, the RCI_{HSU} and

SRB_{CH} or SRB_{MCS} methods can be used by clinicians to determine the significance of reliable changes with relative ease.

One of the strengths of the present study was examining reliable test score changes from multiple perspectives of clinically significant change. As previously noted, reliable test score changes were most strongly associated with diagnostic change, followed by clinicians' ratings, informants' ratings, and to a lesser extent, subjective ratings. Little or no association was expected between subjective reports of loss and reliable deterioration (Hypothesis 13) and partial support was found for this hypothesis. As expected, the correlations between test score changes and reported loss were relatively weak or modest in size. Other studies have also found little to no relation between objectively defined memory deficits and subjective memory loss (Bolla et al., 1991; Feher et al., 1994; McGlone et al., 1990; O'Connor et al., 1990; Sunderland et al., 1986). It has been suggested that individuals may make poor appraisals of their cognitive abilities, either by over-estimating their abilities (e.g., lack of insight or awareness of deficit) or by under-estimating their actual performance. Contrary to expectation, the majority of change scores showed a significant association between subjective reports of cognitive decline and the sum of reliably deteriorated scores. Moreover, subjective ratings of loss were related to loss on some memory measures such as the BCR Retrieval and RAVLT Total but not measures of other cognitive domains. This finding suggests that certain memory measures may be more important than others in classifying the type of change that may be noticed by an individual with declining memory abilities.

The low correlations between reliable change scores and subjective reports may be related, at least in part, to how subjective loss was defined in this study. CSHA

participants were asked "Do you feel that you have more problems with memory than most?" as one item on a self-report measure of mood state. The fact that the question was asked in the context of depression screening may have biased how respondents interpreted and responded to the question. In addition, the question does not address cognitive loss per se and may have been open to a broad range of interpretations by older adults.

As expected (Hypothesis 14), informants' reports of loss were associated with reliable deterioration across all statistical methods when the number of test scores showing reliable change (decline) were summed. The strength of the these correlations were modest but consistently higher than those obtained between test score changes and subjective ratings of loss. As with subjective ratings, memory tests such as the BCR had the strongest and most consistent associations with informants' rating of decline. In addition, informant's ratings of loss were also associated with reliable test score decline on timed measures of word fluency and processing speed. These findings suggest that informants' ratings of loss were more likely to reflect multiple or diffuse, rather than specific, cognitive changes over a 5-year interval.

Among the non-diagnostic indices of clinically significant change, clinicians' rating of loss in cognition was most strongly associated with reliable deterioration in test performance. As expected, clinicians' reports of loss were associated with the sum of reliably deteriorated test scores across all statistical methods (Hypothesis 15). The strength of the these correlations were consistently higher than those obtained between test score changes and informants' ratings of loss. As with ratings made by informants and study participants, memory tests such as the BCR and WMS Information subtest had

the strongest and most consistent associations with clinicians' ratings of decline. In addition, clinicians' ratings were also associated with reliable test score decline on measures of verbal abstraction and semantic fluency using select statistical methods.

To the best of the author's knowledge, no other study has demonstrated that reliably determined change on neuropsychological tests is related to judgments of cognitive loss. The fact that informants' and clinicians' ratings of cognitive change coincided with reliable deterioration in test performance is a step toward validating the use of neuropsychological tests and change score methods for detecting change in older adults. However, it is important to note that reliable change was never perfectly associated with any index of clinical significance and is probably best viewed as one component of determining clinically significant change (Beutler & Moleiro, 2001).

Limitations. This study sought to extend the methods of previous research by examining several change score methods from a variety of perspectives pertaining to clinically significant change. It accomplished this using neuropsychological data from a large national study of older adults with multiple cognitive measures. In spite of these improvements over prior research, there are several limitations to the current study. The first relates to attrition (a threat in most longitudinal studies) and in particular, the small sample size of the group that progressed to dementia. Although the CSHA is a population-based study of older adults across Canada, the small number of individuals who progressed to dementia and completed the neuropsychological assessment at follow-up may not be representative of persons with dementia as a whole. The extent to which the study findings can be generalized may be limited. There is, however, no gross indication of a selection bias given that there were no significant differences in

demographic characteristics or baseline test performance among persons who progressed to dementia and either did or did not complete follow-up testing.

Another limitation in this study was the large number of analyses that were conducted may have increased the probability of Type I error. To minimize this risk, corrections for testwise error were applied to group-level analyses and stringent criteria (i.e., alpha < 0.01) were used to interpret chi-square tests of significance in the individual-level analyses. A feature of the study design, however, was that alpha needed to be set at relatively higher rates (i.e., 0.05 or 0.10) in the calculation of reliable change. As previously mentioned, it is important for researchers who use change score methodology to remember that when the null hypothesis is tested for each individual in a large sample across multiple measures using several different methods, it is possible that the findings will be unduly influenced by Type I error. Consistent with Keith et al's (2002) suggestion, this risk was reduced by considering the utility of the change score methods across several neuropsychological tests in relation to chance levels of misclassification.

An obvious, but unexpected, limitation of this study relates to the examination of the RC_{ID} and RC_{INDIV}. Missing data and low r_{DD}s may have precluded a full and adequate comparison of these methods to other change score methods. As a result, they may have appeared less accurate or useful than they might be in a different sample of older adults or in younger individuals. The problems encountered using these methods in this study might also have been avoided if the study design had employed different measures and/or different test-retest intervals. Future research is needed to re-examine the utility of these change score methods.

As in many longitudinal studies, a limitation of this study involves the use of outdated measures. The measures used in CSHA were initially selected in 1990. Since then, new measures have been developed for use with the general population and with older adults in particular. The third revision of the Wechsler Adult Intelligence Scale makes it less likely that clinicians will continue to employ the subtests from older versions, particularly the short-forms of these tests that were employed in the CSHA. On the other hand, measures such as the BCR may become more popular given the wealth of normative data that are now available from the CSHA and the favourable evidence of its utility in change score measurement with older adults.

Criterion contamination may pose a threat to the internal validity of the study, particularly those findings pertinent to diagnostic change. Diagnostic ratings in the CSHA were at least partially based on neuropsychological test performance. This may have served to inflate the rate of association between diagnostic change and test score change that was found in this study. Though the potential for criterion contamination exists, it should be noted that the final diagnosis in the CSHA was based on more than just neuropsychological test data or a neuropsychologist's opinion – it also involved physicians' judgments based on history and examination, information provided by informants, and a weighing of all relevant data. Physicians did not have access to test score differences when they made their judgments about loss and neuropsychologists did not have access to age-corrected changes in test performance making this threat less likely in a study of test score changes. In this study, criterion contamination is not an issue that would have affected informant's ratings of change.

Future directions. Based on the results of this study, it may be fruitful for future research to examine and validate change score methods, particularly the RCI_{HSU}, in other clinical samples of older adults to determine whether these methods can assist in the early detection of particular neurodegenerative disorders. It will also be necessary to determine the utility of these methods with shorter test-retest intervals, perhaps on the order of 1 or 2 years, as these are more common retest intervals in clinical practice. It may be that change score methods are most useful only after a certain amount of time has elapsed between evaluations.

In clinical settings, it is the progression of deficits and the pattern of changes that occur early in the course of the disorder that help distinguish one form of dementia from another. Using a broader range of tests that assess more cognitive domains (e.g., executive functioning, fine motor speed, working memory) in the future may help to facilitate the accurate identification of dementia or cognitive impairment by increasing the sensitivity of change score methods. The identification of particular patterns of loss in an individual may specifically facilitate the early detection of dementia and contribute to dementia differential diagnosis (e.g., Alzheimer Disease, Lewy Body Dementia, and Frontotemporal Dementia).

In light of advances in medical technology and the greater emphasis that has been placed on multidisciplinary work in clinical settings, future research could also examine the relation between test score changes and other investigative techniques such as changes in CT or MR images and how these can be combined to render accurate prognostic information about a particular individual.

References

Abramson, I. S. (2000). Reliable change formula query: A statistician's comments. Journal of the International Neuropsychological Society, 6, 365.

Albert, M. S. (1994). Age-related changes in cognitive function. In M. L. Albert & J. E. Knoefel (Eds.), Clinical neurology of aging (2nd ed., pp. 314-328). New York: Oxford University Press.

Albert, M. S., Moss, M. B., Tanzi, R., & Jones, K. (2002). Preclinical prediction of AD using neuropsychological tests. Journal of the International Neuropsychological Society, 7, 631-639.

American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders (3rd rev. ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: American Psychiatric Association.

Anastasi, A. (1988). Psychological testing (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Aron, A., & Aron, E. N. (1994). Statistics for psychology. Englewood Cliffs, NJ: Prentice-Hall.

Bancher, C., Paulus, W., Paukner, K., & Jellinger, K. (1997). Neuropathologic diagnosis of Alzheimer disease: Consensus between practicing neuropathologists? Alzheimer Disease and Associated Disorders, 11, 207-219.

Barona, A., & Reynolds, C. R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. Journal of Consulting and Clinical Psychology, 52, 885-887.

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

Benton, A. L. (1974). Revised visual retention test: Clinical and experimental applications. New York: Psychological Corporation.

Benton, A. L., & Hamsher, K. (1978). Multilingual Aphasia Examination manual. Iowa City: University of Iowa Press.

Beutler, L. E., & Moleiro, C. (2001). Clinical versus reliable and significant change. Clinical Psychology: Science and Practice, 8, 441-445.

Bieliauskas, L. A., Fastenau, P. S., Lacy, M. A., & Roper, B. L. (1997). Use of the odds ratio to translate neuropsychological test scores into real-world outcomes: From statistical significance to clinical significance. Journal of Clinical and Experimental Neuropsychology, 19, 889-896.

Bolla, K. I., Lindgren, K. N., Bonaccorsy, C., & Bleeker, M. L. (1991). Memory complaints in older adults: Fact or fiction? Archives of Neurology, 48, 61-64.

Bondi, M. W., Monsch, A. U., Galasko, D., Butters, N., Salmon, D. P., & Delis, D. C. (1994). Preclinical cognitive markers of dementia of the Alzheimer type. Neuropsychology, 8, 374-384.

Brink, T. L., Yesavage, J. A., Lum, O., Heirsema, P., Adey, M., & Rose, T. L. (1982). Screening tests for geriatric depression. Clinical Gerotonologist, 1, 37-43.

Brophy, A. L. (1986). Confidence intervals for true scores and retest scores on clinical tests. Journal of Clinical Psychology, 42, 989-991.

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

Brull, J., Wertheimer, J., & Haller, E. (1979). Evolutive profiles in senile dementia: A psychological and neuropsychological longitudinal study. In F. Hoffmeister & C. Muller (Eds.), Brain function in old age: Evaluation of changes and disorders (pp. 188-205). New York: Springer-Verlag.

Buschke, H. (1984). Cued recall in amnesia. Journal of Clinical Neuropsychology, 6, 433-440.

Cahn, D. A., Sullivan, E. V., Shear, P. K., Heit, G., Lim, K. O., Marsh, L., Lane, B., Wasserstein, P., & Silverberg, G. D. (1998). Neuropsychological and motor functioning after unilateral anatomically guided posterior ventral pallidotomy: Preoperative performance and three-month follow-up. Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 11, 136-145.

Canadian Study of Health and Aging Working Group. (1994). Canadian Study of Health and Aging: Study methods and prevalence of dementia. Canadian Medical Association Journal, 150, 899-913.

Canadian Study of Health and Aging Working Group. (2000). The incidence of dementia in Canada. Neurology, 55, 66-73.

Charter, R. A. (1996). Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. Perceptual and Motor Skills, 82, 1139-1142.

Chelune, G. J. (1998, February). Reliable neuropsychological change: An elaboration of methods for assessing meaningful change. Paper presented at the annual meeting of the International Neuropsychological Society, Honolulu, HI.

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

Chen, S. H., Thomas, J. D., Glueckauf, R. L., & Bracy, O. L. (1997). The effectiveness of computer-assisted cognitive rehabilitation for persons with traumatic brain injury. Brain Injury, 11, 197-209.

Christensen, H., Mackinnon, A. J., Korten, A. E., Jorm, A. F., Henderson, A. S., Jacomb, P., & Rodgers, B. (1999). An analysis of diversity in the cognitive performance of elderly community dwellers: Individual differences in change scores as a function of age. Psychology and Aging, 14, 365-379.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC Index. Behavior Therapy, 17, 305-308.

Clark, C. M., Sheppard, L., Fillenbaum, G. G., Galasko, D., Morris, J. C., Koss, E., Mohs, R., & Heyman, A. (1999). Variability in annual Mini-Mental State Examination score in patients with probable Alzheimer disease: A clinical perspective of data from the Consortium to Establish a Registry for Alzheimer's Disease. Archives of Neurology, 56, 857-862.

Colsher, P. L., & Wallace, R. B. (1991). Longitudinal application of cognitive function measures in a defined population of community-dwelling elders. Annals of Epidemiology, 1, 215-230.

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. Journal of Clinical and Experimental Neuropsychology, 20, 755-762.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" - or should we? Psychological Bulletin, 74, 68-80.

Cronbach, L. J., & Gleser, G. C. (1959). Interpretation of reliability and validity coefficients: Remarks on a paper by Lord. Journal of Educational Psychology, 50, 230-237.

Crystal, H. A., Dickson, D., Sliwinski, M., Masur, D., Blau, A., & Lipton, R. B. (1996). Associations of status and change measures of neuropsychological function with pathologic changes in elderly, originally nondemented subjects. Archives of Neurology, 53, 82-87.

Cummings, J. L., & Benson, D. F. (1983). Dementia: A clinical approach. Boston, MA: Butterworth.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). California Verbal Learning Test. San Antonio, TX: Psychological Corporation.

Dikmen, S., Machamer, J., Temkin, N., & McLean, A. (1990). Neuropsychological recovery in patients with moderate to severe head injury: 2 year follow-up. Journal of Clinical and Experimental Neuropsychology, 12, 507-519.

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

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. Psychological Bulletin, 86, 335-337.

Dupont, W. D., & Plummer, W. D. (1997). Power and Sample Size Calculation (Version 1.0.15) [Computer program]. Nashville, TN: Vanderbilt University Medical Center.

Edwards, D. W., Yarvis, R. M., Mueller, D. P., Zingale, H. C., & Wagman, W. J. (1978). Test-taking and the stability of adjustment scales. Can we assess patient deterioration? Evaluation Quarterly, 2, 275-291.

Essex-Sorlie, D. (1995). Medical biostatistics and epidemiology. Norwalk, CT: Appleton & Lange.

Feher, E. P., Larrabee, G. J., Sudilovsky, A., & Crook, T. H. (1994). Memory self-report in Alzheimer's disease and in age-associated memory impairment. Journal of Geriatric Psychiatry and Neurology, 7, 58-65.

Flicker, C., Ferris, S. H., Crook, T., Bartus, R. T., & Reisberg, B. (1985). Cognitive function in normal aging and early dementia. In J. Traber & W. H. Gispen (Eds.), Senile dementia of the Alzheimer type: Early diagnosis, animal pathology, and animal models (pp. 2-17). New York: Springer-Verlag.

Flicker, C., Ferris, S. H., & Reisberg, B. (1993). A two-year longitudinal study of cognitive function in normal aging and Alzheimer's disease. Journal of Geriatric Psychiatry and Neurology, 6, 84-96.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189-198.

Frank, R., Wiederholt, W. C., Kritz-Silverstein, D., Salmon, D. P., & Barrett-Connor, E. (1996). Effects of sequential neuropsychological testing of an elderly community-based sample. Neuroepidemiology, 15, 257-268.

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits for test scores. Educational and Psychological Measurement, 47, 607-614.

Gottman, J. M., & Rushe, R. H. (1993). The analysis of change: Issues, fallacies, and new ideas. Journal of Consulting and Clinical Psychology, 61, 907-910.

Graves, R. E. (2000). Accuracy of regression equation prediction across the range of estimated premorbid IQ. Journal of Clinical and Experimental Neuropsychology, 22, 316-324.

Graves, R. E., Carswell, L. M., & Snow, W. G. (1999). An evaluation of the sensitivity of premorbid IQ estimators for detecting cognitive decline. Psychological Assessment, 11, 29-38.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons.

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10, 255-282.

Hageman, W. J. J. M., & Arrindell, W. A. (1993). A further refinement of the reliable change (RC) index by improving the pre-post difference score: Introducing RC(ID). Behaviour Research and Therapy, 31, 693-700.

Hageman, W. J. J. M., & Arrindell, W. A. (1999a). Clinically significant and practical! Enhancing precision does make a difference. Reply to McGlinchey and Jacobson, Hsu, and Speer. Behaviour Research and Therapy, 37, 1219-1233.

Hageman, W. J. J. M., & Arrindell, W. A. (1999b). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. Behaviour Research and Therapy, 37, 1169-1193.

Harris, C. W. (Ed.). (1963). Problems measuring change. Madison, WI: University of Wisconsin Press.

Heaton, R. (1999). Detecting change: A comparison of three methods using normal and clinical samples. Paper presented at the mid-annual meeting of the International Neuropsychological Society, Durban, South Africa.

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

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., & Grant, I. (2001). A comparison of three neuropsychological methods, using normal and clinical samples. Archives of Clinical Neuropsychology, 16, 75-91.

Helmstaedter, C., Gleibner, U., Zentner, J., & Elger, C. E. (1998). Neuropsychological consequences of epilepsy surgery in frontal lobe epilepsy. Neuropsychologia, 36, 333-341.

Hermann, B. P., Seidenberg, M., Schoenfeld, J., Peterson, J., Leveroni, C., & Wyler, A. R. (1996). Empirical techniques for determining the reliability, magnitude, and

pattern of neuropsychological change after epilepsy surgery. Epilepsia, 37, 942-950.

Hermann, B. P., & Wyler, A. R. (1988). Neuropsychological outcome of anterior temporal lobectomy. Journal of Epilepsy, 1, 35-45.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1992). Intra-individual change in text recall of the elderly. Brain and Language, 42, 248-269.

Hinton-Bayre, A. (2000). Reliable change formula query. Journal of the International Neuropsychological Society, 6, 362-363.

Hinton-Bayre, A. D., Geffen, G. M., Geffen, L. B., McFarland, K. A., & Friis, P. (1999). Concussion in contact sports: Reliable change indices of impairment and recovery. Journal of Clinical and Experimental Neuropsychology, 21, 70-86.

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

Howieson, D. B., Dame, A., Camicioli, R., Sexton, G., Payami, H., & Kaye, J. A. (1997). Cognitive markers preceding Alzheimer's dementia in the healthy oldest old. Journal of the American Geriatrics Society, 45, 584-589.

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. Behavioral Assessment, 11, 459-467.

Hsu, L. M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. Journal of Consulting and Clinical Psychology, 63, 141-144.

Hsu, L. M. (1999). A comparison of three methods of identifying reliable and clinically significant client changes: Commentary on Hageman and Arrindell. Behaviour Research and Therapy, 37, 1195-1202.

Iverson, G. (1999). Interpreting change on the WAIS-III/WMS-III following traumatic brain injury. Journal of Cognitive Rehabilitation, 17, 16-20.

Iverson, G. (2000). Interpreting change on the WAIS-III/WMS-III in clinical samples. Archives of Clinical Neuropsychology, 16, 183-191.

Iverson, G. L., & Green, P. (2001). Measuring improvement or decline on the WAIS-R in inpatient psychiatry. Psychological Reports, 89, 457-462.

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

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's Older Americans Normative Studies: WAIS-R, WMS-R, and AVLT norms for ages 56 through 97. The Clinical Neuropsychologist, 6 (Suppl.), 1-104.

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

Ivnik, R. J., Smith, G. E., Lucas, J. A., Tangalos, E. G., Kokmen, E., & Petersen, R. C. (1997). Free and cued selective reminding test: MOANS norms. Journal of Clinical and Experimental Neuropsychology, 19, 676-691.

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

Jacobs, D. M., Sano, M., Dooneief, G., Marder, K., Bell, K. L., & Stern, Y. (1995). Neuropsychological detection and characterization of preclinical Alzheimer's disease. Neurology, 45, 957-962.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. Behavior Therapy, 15, 336-352.

Jacobson, N. S., & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. Behavioral Assessment, 10, 133-145.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. Journal of Consulting and Clinical Psychology, 67, 300-307.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59, 12-19.

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). The Boston Naming Test (2nd ed.). Philadelphia: Lea & Febiger.

Kazdin, A. (1999). The meanings and measurement of clinical significance. Journal of Consulting and Clinical Psychology, 67, 332-339.

Keith, J. R., Puente, A. E., Malcolmson, K. L., Tartt, S., Coleman, A. E., & Marks, H. F. (2002). Assessing postoperative cognitive change after cardiopulmonary bypass surgery. Neuropsychology, 16, 411-421.

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

Korten, A. E., Henderson, A. S., Christensen, H., Jorm, A. F., Rodgers, B., Jacomb, P., & Mackinnon, A. J. (1997). A prospective study of cognitive function in the elderly. Psychological Medicine, 27, 919-930.

Lezak, M. D. (1995). Neuropsychological assessment (3rd ed.). New York: Oxford University Press.

Libon, D. J., Mattson, R. E., Glosser, G., Kaplan, E., Malamut, B. M., Sands, L. P., Swenson, R., & Cloud, B. S. (1996). A nine-word dementia version of the California Verbal Learning Test. The Clinical Neuropsychologist, 10, 237-244.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 47, 121-150.

Lord, F. M. (1957). The measurement of growth. Educational and Psychological Measurement, 16, 421-437.

Lord, F. M. (1958). Further problems in the measurement of growth. Educational and Psychological Measurement, 18, 437-451.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change (pp. 21-38). Madison, WI: University of Wisconsin Press.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lucas, J. A., Ivnik, R. J., Smith, G. E., Bohac, D. L., Tangalos, E. G., Graff-Radford, N. R., & Petersen, R. C. (1998). Mayo's Older Americans Normative Studies: Category fluency norms. Journal of Clinical and Experimental Neuropsychology, 20, 194-200.

Lucas, J. A., Ivnik, R. J., Smith, G. E., Bohac, D. L., Tangalos, E. G., Kokmen, E., Graff-Radford, N. R., & Petersen, R. C. (1998). Normative data for the Mattis Dementia Rating Scale. Journal of Clinical and Experimental Neuropsychology, 20, 536-547.

Maassen, G. H. (2000a). Kelley's formula as a basis for the assessment of reliable change. Psychometrika, 65, 187-197.

Maassen, G. H. (2000b). Principles of defining reliable change indices. Journal of Clinical and Experimental Neuropsychology, 22, 622-632.

Maassen, G. H. (2001). The unreliable change of reliable change indices. Behaviour Research and Therapy, 39, 495-498.

Mahanna, E. P., Blumenthal, J. A., White, W. D., Croughwell, N. D., Clancy, C. P., Smith, L. R., & Newman, M. F. (1996). Defining neuropsychological dysfunction after coronary artery bypass grafting. Annals of Thoracic Surgery, 61, 1342-1347.

Masur, D. M., Sliwinski, M., Lipton, R. B., Blau, A. D., & Crystal, H. A. (1994). Neuropsychological prediction of dementia and the absence of dementia in healthy elderly persons. Neurology, 44, 1427-1432.

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

Mattis, S. (1988). Dementia Rating Scale (DRS): Professional manual. Odessa, FL: Psychological Assessment Resources.

McCaffrey, R. J., Ortega, A., & Haase, R. F. (1993). Effects of repeated neuropsychological assessments. Archives of Clinical Neuropsychology, 8, 519-524.

McCaffrey, R. J., Ortega, A., Orsillo, S. M., Nelles, W. B., & Haase, R. F. (1992). Practice effects in repeated neuropsychological assessments. The Clinical Neuropsychologist, 6, 32-42.

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

McCleary, R., Dick, M. B., Buckwalter, G., Henderson, V., & Rodman Shankle, W. (1996). Full-information models for multiple psychometric tests: Annualized rates of change in normal aging and dementia. Alzheimer Disease and Associated Disorders, 10, 216-223.

McGlinchey, J. B., & Jacobson, N. S. (1999). Clinically significant but impractical? A response to Hageman and Arrindell. Behaviour Research and Therapy, 37, 1211-1217.

McGlone, J., Gupta, S., Humphrey, D., Oppenheimer, S., Mirsen, T., & Evans, D. R. (1990). Screening for early dementia using memory complaints from patients and relatives. Archives of Neurology, 47, 1189-1193.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA workgroup under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. Neurology, 34, 939-944.

McNemar, Q. (1958). On growth measurement. Educational and Psychological Measurement, 18, 47-55.

McSweeney, A. J., Naugle, R. I., Chelune, G. J., & Luders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. The Clinical Neuropsychologist, 7, 300-312.

Medical Research Council of Canada, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (1998, August). Tri-Council policy statement: Ethical conduct for research involving humans. Retrieved October 3, 2001, from the World Wide Web: http://www.nserc.ca/programs/ethics/english/ethics-e.pdf

Mega, M. S., Masterman, D. L., Benson, F., Vinters, H. V., Tomiyasu, U., Craig, A. H., Foti, D. J., Kaufer, D., Scharre, D. W., Fairbanks, L., & Cummings, J. L. (1996). Dementia with Lewy bodies: Reliability and validity of clinical and pathologic criteria. Neurology, 47, 1403-1409.

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

Morris, J., Heyman, A., & Mohs, R. (1989). The consortium to establish a registry for Alzheimer's disease (CERAD). Part I: Clinical and neuropsychological assessment of Alzheimer's disease. Neurology, 39, 1159-1165.

Morris, J. C., Edland, S., Clark, C., Galasko, D., Koss, E., Mohs, R., van Belle, G., Fillenbaum, G., & Heyman, A. (1993). The consortium to establish a registry for Alzheimer's disease (CERAD). Part IV. Rates of cognitive change in the longitudinal assessment of probable Alzheimer's disease. Neurology, 43, 2457-2465.

Nagy, Z., Esiri, M. M., Joachim, C., Jobst, K. A., Morris, J. H., King, E. M. F., Hindley, N. J., McDonald, B., Litchfield, S., Barnetson, L., & Smith, A. D. (1998). Comparison of pathological diagnostic criteria for Alzheimer disease. Alzheimer Disease and Associated Disorders, 12, 182-189.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. Psychological Bulletin, 88, 622-637.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw Hill.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw Hill.

Nunnally, J. C., & Kotsch, W. E. (1983). Studies of individual subjects: Logic and methods of analysis. British Journal of Clinical Psychology, 22, 83-93.

O'Connor, D. W., Pollitt, P. A., Roth, M., Brook, C. P. B., & Reiss, B. B. (1990). Memory complaints and impairment in normal, depressed, and demented elderly

persons identified in a community survey. Archives of General Psychiatry, 47, 224-227.

Paolo, A. M., & Ryan, J. J. (1992). Generalizability of two methods of estimating premorbid intelligence in the elderly. Archives of Clinical Neuropsychology, 7, 135-143.

Paolo, A. M., & Ryan, J. J. (1993). Test-retest stability of the Satz-Mogel WAIS-R short form in a sample of normal persons 75 to 87 years of age. Archives of Clinical Neuropsychology, 8, 397-404.

Paolo, A. M., Troster, A. I., & Ryan, J. J. (1997). California Verbal Learning Test: Normative data for the elderly. Journal of Clinical and Experimental Neuropsychology, 19, 220-234.

Payne, R. W., & Jones, H. G. (1957). Statistics for the investigation of individual cases. Journal of Clinical Psychology, 13, 115-121.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Phillips, N. A., & McGlone, J. (1995). Grouped data do not tell the whole story: Individual analysis of cognitive change after temporal lobectomy. Journal of Clinical and Experimental Neuropsychology, 17, 713-724.

Purdon, S. E., Jones, B. D. W., Stip, E., Labelle, A., Addington, D., David, S. R., Breier, A., & Tollefson, G. D. (2000). Neuropsychological change in early phase schizophrenia during 12 months of treatment with olanzapine, risperidone, or haloperidol. Archives of General Psychiatry, 57, 249-258.

Reitan, R., & Wolfson, D. (1985). The Halstead-Reitan Neuropsychological Test Battery. Tempe, AZ: Neuropsychology Press.

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

Rey, A. (1964). L'examen clinique en psychologie. Paris: Presses Universitaires de France.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie & R. T. Campbell & W. Meredith & S. C. Rawlings (Eds.), Methodological issues in aging research (pp. 171-209). New York: Springer.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726-748.

Roth, M., Huppert, F. A., Tym, E., & Mountjoy, C. Q. (1988). The Cambridge examination for mental disorders of the elderly. Cambridge: Cambridge University Press.

Rubin, E. H., Storandt, M., Philip Miller, J., Kinscherf, D. A., Grant, E. A., Morris, J. C., & Berg, L. (1998). A prospective study of cognitive function and onset of dementia in cognitively healthy elders. Archives of Neurology, 55, 395-401.

Ryan, J. J., Paolo, A. M., & Brungardt, T. M. (1990). Standardization of the Wechsler Adult Intelligence Scale - Revised for persons 75 years and older. Psychological Assessment, 2, 404-411.

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

Sandercock, P. (1989). The odds ratio: A useful tool in neurosciences. Journal of Neurology, Neurosurgery, and Psychiatry, 52, 817-820.

Satz, P., & Mogel, S. (1962). An abbreviation of the WAIS for clinical use. Journal of Clinical Psychology, 18, 77-79.

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

Sawrie, S. M., Marson, D. C., Boothe, A. L., & Harrell, L. E. (1999). A method for assessing clinically relevant individual cognitive change in older adult populations. Journal of Gerontology: Psychological Sciences, 54B, P116-P124.

Schaie, K. W. (1989). Perceptual speed in adulthood: Cross-sectional and longitudinal studies. Psychology and Aging, 4, 443-453.

Schaie, K. W. (1996). Intellectual development in adulthood: The Seattle Longitudinal Study. New York: Cambridge University Press.

Scherzer, B. P. (1986). Rehabilitation following severe head trauma: Results of a three-year program. Archives of Physical Medicine and Rehabilitation, 67, 366-374.

Schinka, J. A., & Vanderploeg, R. D. (2000). Estimating premorbid level of functioning. In R. D. Vanderploeg (Ed.), Clinician's guide to neuropsychological assessment (2nd ed., pp. 39-67). Mahwah, NJ: Lawrence Erlbaum Associates.

Shammi, P., Bosman, E., & Stuss, D. T. (1998). Aging and variability in performance. Aging, Neuropsychology, and Cognition, 5, 1-13.

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

Shaw, P. J., Bates, D., Cartlidge, N. E. F., French, J. M., Heaviside, D., Julian, D. G., & Shaw, D. A. (1986). Early intellectual dysfunction following coronary bypass surgery. Quarterly Journal of Medicine, 225, 59-68.

Smith, A. (1982). Symbol Digit Modalities Test manual - revised. Los Angeles: Western Psychological Services.

Sohlberg, M. M., & Mateer, C. A. (1989a). Introduction to cognitive rehabilitation: Theory and practice. New York: Guildford Press.

Sohlberg, M. M., & Mateer, C. A. (1989b). Training use of compensatory memory books: A three stage behavioral approach. Journal of Clinical and Experimental Neuropsychology, 11, 871-891.

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. Journal of Consulting and Clinical Psychology, 60, 402-408.

Speer, D. C. (1999). What is the role of two-wave designs in clinical research? Comment on Hageman and Arrindell. Behaviour Research and Therapy, 37, 1203-1210.

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. Journal of Consulting and Clinical Psychology, 63, 1044-1048.

Spreen, O., & Strauss, E. (1998). A compendium of neuropsychological tests: Administration, norms, and commentary (2nd ed.). New York: Oxford University Press.

SPSS Inc. (1998). SPSS for Windows (Version 10.0) [Computer software]. Chicago, IL: Author.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 356-442). Washington, DC: American Council on Education.

Statistics Canada. (1991). Population projections 1990-2011 based on recent changes in fertility levels and revised immigration targets. Ottawa, ON: Author.

Statistics Canada. (2002). Population by age and sex. Retrieved December 5, 2002, from the World Wide Web: http://www.statcan.ca/english/Pgdb/demo10a.htm

Storandt, M., Botwinick, J., & Danziger, W. L. (1986). Longitudinal changes: Patients with mild SDAT and matched healthy controls. In L. W. Poon (Ed.), Clinical memory assessment of older adults (pp. 277-284). Washington, DC: American Psychological Association.

Streiner, D. L., & Norman, G. R. (1995). Health measurement scales: A practical guide to their development and use (2nd ed.). New York: Oxford University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643-662.

Sunderland, A., Watts, J., Baddeley, A. D., & Harris, J. E. (1986). Subjective memory assessment and test performance in elderly adults. Journal of Gerontology, 41, 376-384.

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics (3rd ed.). New York, NY: Harper Collins.

Taylor, R. (1998). Indices of neuropsychological functioning and decline over time in dementia. Archives of Gerontology and Geriatrics, 27, 165-170.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. Journal of the International Neuropsychological Society, 5, 357-369.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (2000). Reliable change formula query: Temkin et al. reply. Journal of the International Neuropsychological Society, 6, 364.

Teng, E. L., & Chui, H. C. (1987). The modified mini-mental state (3MS) examination. Journal of Clinical Psychiatry, 48, 314-318.

Theisen, M. E., Rapport, L. J., Axelrod, B. N., & Brines, D. B. (1998). Effects of practice in repeated administrations of the Wechsler Memory Scale - Revised in normal adults. Assessment, 5, 85-92.

Tiffin, J. (1968). Purdue Pegboard examiner's manual. Rosemont, IL: London House.

Tombaugh, T. N., & Hubley, A. M. (1997). The 60-item Boston Naming Test: Norms for cognitively intact adults aged 25 to 88 years. Journal of Clinical and Experimental Neuropsychology, 19, 922-932.

Tombaugh, T. N., & McIntyre, N. J. (1992). The Mini-Mental State Examination: A comprehensive review. Journal of the American Geriatrics Society, 40, 922-935.

Tuokko, H., & Hadjistavropoulos, T. (1998). An assessment guide to geriatric neuropsychology. Mahwah, NJ: Lawrence Erlbaum Associates.

Tuokko, H., Hadjistavropoulos, T., Miller, J. A., Horton, A., & Beattie, B. L. (1995). The Clock Test: Administration and scoring manual. Toronto, ON: Multi-Health Systems.

Tuokko, H., Kristjansson, E., & Miller, J. A. (1995). The neuropsychological detection of dementia: An overview of the neuropsychological component of the Canadian Study of Health and Aging. Journal of Clinical and Experimental Neuropsychology, 17, 352-373.

Tuokko, H., & Woodward, T. S. (1996). Development and validation of a demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. Journal of Clinical and Experimental Neuropsychology, 18, 479-616.

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

Wechsler, D. (1975). Wechsler Memory Scale. New York: Psychological Corporation.

Wechsler, D. (1981). WAIS-R manual. New York: Psychological Corporation.

Wechsler, D. (1987). Wechsler Memory Scale - Revised manual. San Antonio, TX: Psychological Corporation.

Weinstein, A., Schwid, S. I., Schiffer, R. B., McDermott, M. P., Giang, D. W., & Goodman, A. D. (1999). Neuropsychologic status in multiple sclerosis after treatment with glatiramer. Archives of Neurology, 56, 319-324.

Wilson, B. A., Watson, P. C., Baddeley, A. D., Emslie, H., & Evans, J. J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. Journal of the International Neuropsychological Society, 6, 469-479.

World Health Organization. (1987). Tenth revision of the International Classification of Diseases 1987 draft of chapter V, categories F00-F99, mental, behavioural and

developmental disorders, Clinical Descriptions and Diagnostic Guidelines. Geneva: Author.

Zelinski, E. M., & Burnight, K. P. (1997). Sixteen-year longitudinal and time lag changes in memory and cognition in older adults. Psychology and Aging, 12, 503-513.

Table 1

Descriptive data for NCI persons at CSHA-1 who completed neuropsychological testing (n = 575)

Variable	Mean (SD)	Range
Age (years)	79.28 (6.73)	65-98
Education (years)	10.32 (3.81)	0-25
3MS score	83.70 (10.24)	53-100
MMSE score	25.81 (3.16)	14-30
Estimated IQ	100.77 (9.82)	76-121
ADL rating	26.78 (2.14)	10-28
		Percentage
Gender	223 male	38.8
	352 female	61.2
Language	575 English-speaking	100.0
Handedness	537 right	93.4
	31 left / ambidextrous	5.4
	7 missing	1.2
Marital status	32 never married	5.6
	211 married / common-law	36.7
	14 divorced / separated	2.4
	220 widowed	38.3
	98 missing	17.0
Race	566 Caucasian	98.4
	9 other	1.6
Residence	481 community-dwelling	83.7
	94 living in an institution	16.3
Region	172 Atlantic provinces	29.9
	135 Ontario/Quebec	23.5
	116 Prairies	20.2
	152 British Columbia	26.4
Health	150 very good	26.1
	221 pretty good	38.4
	88 not too good	15.9
	11 poor	1.9
	4 very poor	0.7
	101 missing	17.6

Note. NCI = no cognitive impairment. CSHA = Canadian Study of Health and Aging. 3MS = Modified Mini-Mental State Examination. MMSE = Mini-Mental State Examination. ADL = Activities of Daily Living.

Table 2

Five-year follow-up data for NCI persons at CSHA-1 who completed neuropsychological testing (n = 575)

Variable		Percentage
Participation status	345 participated	60.0
	49 refused participation	8.5
	21 lost to follow-up	3.7
	160 dead	27.8
Participants seen at CSHA-2 (n = 345)		
Consensus diagnosis	208 NCI	60.3
	88 CIND	25.5
	49 Dementia	14.2
Subjective rating of decline	268 no	77.7
	29 yes	8.4
	48 missing	13.9
Informant rating of decline	210 no	60.9
	106 yes	30.7
	29 missing	8.4
Clinician rating of decline	207 no	60.0
	59 yes	17.1
	79 missing	22.9

Note. CSHA = Canadian Study of Health and Aging. NCI = no cognitive impairment. CIND = cognitive impairment, no dementia

Table 3

Descriptive data for persons with NCI at both CSHA-1 and CSHA-2 (n = 166)

Variable	Mean (SD)	Range
Age (years)	76.86 (5.78)	65-93
Education (years)	11.32 (3.91)	0-25
3MS score at CSHA-1	88.17 (9.32)	58-100
MMSE score at CSHA-1	27.06 (2.65)	17-30
Estimated IQ	103.26 (9.70)	83-121
ADL rating at CSHA-1	27.39 (1.26)	22-28
Re-test interval (months)	58.16 (4.50)	42 - 70
		Percentage
Gender	71 male	42.8
	95 female	57.2
Language	166 English-speaking	100.0
Handedness	152 right	91.6
	10 left / ambidextrous	6.0
	4 missing	2.4
Marital status	12 never married	7.2
	75 married / common-law	44.6
	6 divorced / separated	3.6
	59 widowed	35.5
	14 missing	8.4
Race	163 Caucasian	98.2
	3 other	1.8
Residence	152 community-dwelling	91.6
	14 living in an institution	8.4
Region	29 Atlantic provinces	17.5
	40 Ontario/Quebec	24.1
	38 Prairies	22.9
	59 British Columbia	35.5
Health	53 very good	31.9
	76 pretty good	45.8
	17 not too good	10.2
	4 poor	2.4
	0 very poor	0.0
	16 missing	9.6

Note. NCI = no cognitive impairment. CSHA = Canadian Study of Health and Aging. 3MS = Modified Mini-Mental State Examination. MMSE = Mini-Mental State Examination. ADL = Activities of Daily Living.

Table 4

Descriptive data for NCI participants at CSHA-1 who had dementia by CSHA-2 (n = 20)

Variable	Mean (SD)	Range
Age (years)	81.00 (5.62)	67-89
Education (years)	10.00 (4.10)	3-17
3MS score at CSHA-1	81.68 (8.36)	68-94
MMSE score at CSHA-1	25.73 (2.74)	21-29
3MS score at CSHA-2	66.57 (14.47)	42-91
MMSE score at CSHA-2	20.50 (4.90)	11-28
Estimated IQ	100.06 (11.33)	86-119
ADL rating at CSHA-1	27.07 (1.44)	23-28
ADL rating at CSHA-2	23.30 (6.22)	9-28
Re-test interval (months)	57.5 (3.76)	52-68
		Percentage
Gender	7 male	35.0
	13 female	65.0
Language	20 English-speaking	100.0
Handedness	19 right	95.0
	1 ambidextrous	5.0
Marital status	1 never married	5.0
	5 married / common-law	25.0
	9 widowed	45.0
	5 missing	25.0
Race	20 Caucasian	100.0
Residence at CSHA-1	15 community-dwelling	75.0
	5 living in an institution	25.0
Residence at CSHA-2	11 community-dwelling	55.0
	9 living in an institution	45.0
Region	9 Atlantic provinces	45.0
	4 Ontario	20.0
	4 Prairies	20.0
	3 British Columbia	15.0
Health	6 very good	30.0
	8 pretty good	40.0
	1 not too good	5.0
	0 poor	0.0
	0 very poor	0.0
	5 missing	25.0

Note. NCI = no cognitive impairment. CSHA = Canadian Study of Health and Aging. 3MS = Modified Mini-Mental State Examination. MMSE = Mini-Mental State Examination. ADL = Activities of Daily Living.

Table 5

Descriptive data for CSHA-3 NCI participants (n = 30)

Variable	Mean (SD)	Range
Age (years)	84.7 (6.10)	75 - 97
Education (years)	12.37 (4.51)	4 - 25
3MS score	89.13 (4.90)	77 - 99
Re-test interval (months)	3.20 (1.21)	1.2 – 7.3
		Percentage
Gender	18 male	60.0
	12 female	40.0
Language	30 English-speaking	100.0
Handedness	29 right	96.7
	1 missing	3.3

Note. CSHA = Canadian Study of Health and Aging. NCI = no cognitive impairment. 3MS = Modified Mini-Mental State Examination.

Table 6

Neuropsychological test data for persons with NCI at CSHA-3 and follow-up (n = 30)

Measures	N	Time 1 Mean (SD)	Time 2 Mean (SD)	Mean difference (SD)	Correlation
Memory					
WMS Information	30	4.60 (1.13)	4.77 (1.04)	0.17 (0.91)	0.65[a] ***
BCR Trial 1 Free Recall	30	7.90 (1.35)	8.70 (1.53)	0.80 (1.56)**	0.42 *
BCR Retrieval	30	26.77 (3.38)	28.43 (3.44)	1.67 (3.18)**	0.57 **
BCR Acquisition	30	35.83 (0.46)	36.00 (0.00)	0.17 (0.46)	-
BCR Delayed Free Recall	30	9.97 (1.16)	10.50 (1.43)	0.53 (1.48)	0.36 *
Abstract thinking					
WAIS-R Similarities	29	10.60 (2.51)	10.31 (2.95)	-0.24 (2.06)	0.68[a] ***
Language					
FAS	29	32.23 (14.05)	33.93 (12.08)	1.79 (6.68)	0.89 ***
Animals	29	15.90 (3.60)	16.34 (3.79)	0.62 (2.97)	0.67 ***
Construction/speed					
WAIS-R Block Design	28	10.69 (4.64)	12.04 (4.49)	1.04 (3.79)	0.65 ***
WAIS-R Digit Symbol	27	33.68 (8.47)	34.67 (9.07)	0.56 (4.54)	0.87 ***

Note. NCI = no cognitive impairment. CSHA = Canadian Study of Health and Aging. All correlations are Pearson product-moment correlation coefficients (based on a normal distribution) with the exception of those Spearman rho correlation coefficients marked with a superscript [a].

* $p < .05$

** $p < .01$

*** $p < .001$

Table 7

Neuropsychological test data for persons with NCI at both CSHA-1 and CSHA-2 (n = 166)

Measures	N	Time 1 Mean (SD)	Time 2 Mean (SD)	Mean difference (SD)	Correlation
Memory					
WMS Information	166	5.42 (0.85)	5.05 (1.21)	-0.37 (1.06)***	0.33[a]***
BCR Trial 1 FR	162	8.35 (1.51)	7.85 (1.69)	-0.49 (1.71)***	0.43***
BCR Retrieval	162	27.91 (3.69)	26.22 (4.29)	-1.69 (3.65)***	0.59***
BCR Acquisition	162	35.81 (0.83)	35.77 (0.95)	-0.05 (1.20)	0.03[a]
BCR Delayed FR	162	10.49 (1.38)	10.15 (1.62)	-0.35 (1.48)**	0.49[a]***
RAVLT Trial 1 FR	153	4.62 (1.74)	4.46 (1.60)	-0.16 (1.95)	0.32***
RAVLT Total Recall	146	41.17 (9.47)	38.58 (9.40)	-2.59 (8.32)***	0.61***
RAVLT Delayed FR	146	8.37 (3.04)	7.38 (3.38)	-0.99 (3.60)**	0.38***
BVRT	156	12.29 (2.34)	12.19 (2.08)	-0.11 (2.31)	0.46***
Abstract thinking					
WAIS-R Similarities	165	8.29 (4.23)	8.41 (4.10)	0.12 (2.99)	0.74***
Judgment					
WAIS-R Comprehension	165	9.96 (3.03)	9.88 (3.10)	-0.07 (2.63)	0.63***
Language					
Token Test	158	39.91 (5.66)	39.43 (5.33)	-0.48 (4.68)	0.47[a]***
FAS	161	32.32 (12.50)	31.53 (12.23)	-0.79 (7.97)	0.79***
Animals	163	15.90 (4.38)	15.15 (4.36)	-0.75 (3.94)*	0.59***
Construction/Speed					
WAIS-R Block Design	160	11.80 (4.63)	10.06 (4.01)	-1.75 (4.17)***	0.54***
WAIS-R Digit Symbol	153	34.97 (10.51)	30.76 (11.54)	-4.20 (6.05)***	0.85***

Note. NCI = no cognitive impairment. All correlations are Pearson product-moment correlation coefficients (based on a normal distribution) with the exception of those Spearman rho correlation coefficients marked with a superscript [a].
* $p < .05$; ** $p < .01$; *** $p < .001$

Table 8

Summary of Phase A findings

Equation	Classification accuracy (chi-square test)	% of sample with reliable deterioration in memory + other cognitive area(s)	% of sample with no reliably deteriorated scores
2 (SD)	2/10	0.0	56.6
3 (RCI)	4/10	0.0	53.3
4 (RCI_{SPEER})	3/10	3.3	70.0
5a ($RCI_{JT}a$)	6/10	0.0	70.0
5b ($RCI_{JT}b$)	4/10	0.0	50.0
5c ($RCI_{JT}c$)	6/10	0.0	70.0
6a ($RCI_{CHEL}a$)	6/10	0.0	63.3
6b ($RCI_{CHEL}b$)	3/10	3.3	33.3
6c ($RCI_{CHEL}c$)	5/10	0.0	63.3
7 (RCI_{HSU})	7/10	3.3	63.3
8 (RC_{ID})	0/10	0.0	100.0
9 (RC_{INDIV})	0/10	0.0	96.7
10 (SRB_{MCS})	3/10	0.0	63.3
11 (SRB_{MULT})	3/10	0.0	66.7
12 (SRB_{CH})	3/10	0.0	70.0

Note. Classification accuracy in this table refers to the number of measures (out of 10) that conform to the expected distribution of change classification.

Table 9

Magnitude of prediction intervals for individuals with NCI over the shorter test-retest interval

Test	Equations									
	2	3&4	5a&6a	5b&6b	5c&6c	7	8	9	10	11
WMS Information	±1.13	±1.11	±1.57	±1.50	±1.50	±1.42	±1.82	-	±1.32	±1.24
BCR Trial 1	±1.34	±1.69	±2.39	±2.56	±2.57	±2.01	±2.65	±0.88	±2.33	±2.41
BCR Retrieval	±3.38	±3.66	±5.18	±5.23	±5.23	±4.58	±4.23	±2.62	±4.75	±5.01
BCR Acquisition	±0.46	-	-	-	±0.76	-	-	-	-	-
BCR Delay recall	±1.16	±1.52	±2.15	±2.42	±2.43	±1.78	±2.48	±1.01	±2.23	±2.27
Similarities	±2.51	±2.16	±3.05	±3.33	±3.40	±2.84	±3.17	±1.78	±3.39	±3.54
FAS	±14.05	±7.85	±11.10	±10.35	±10.99	±10.77	±10.37	±3.50	±9.43	±9.96
Animals	±3.60	±3.39	±4.79	±4.92	±4.88	±4.38	-	±1.11	±4.69	±4.68
Block Design	±4.74	±4.64	±6.56	±6.38	±6.23	±5.95	±6.63	-	±5.74	±5.48
Digit Symbol	±8.47	±5.08	±7.19	±7.45	±7.47	±6.95	-	±2.07	±7.58	±7.77
Average ranking	3	2	7	8	9	4	10	1	5	6

Note. NCI = no cognitive impairment. Prediction intervals are ranked from smallest (1) to largest (10). It was not possible to calculate fixed prediction intervals for equation 12 (SRB_{CH}) as each prediction interval using this method is variable and dependent on the individual's pretest scores. Prediction intervals could not be calculated for BCR Acquisition using several methods due to lack of variability in test scores. Prediction intervals could not be computed for Animals and WAIS-R Digit Symbol subtests using Equation 8 (RC_{ID}) based on inability to calculate Guttman's reliability coefficients for these measures. Prediction intervals could not be calculated for WAIS-R Block Design subtest using Equation 9 (RC_{INDIV}) given a negative r_{DD}.

Table 10

Summary of Phase B findings

Equation	Classification accuracy (goodness of fit)	% of sample with reliable deterioration in memory + other cognitive area	% of sample with no reliably deteriorated scores
2 (SD)	3/16	41.9	19.0
3 (RCI)	3/16	40.2	22.3
4 (RCI_{SPEER})	1/16	25.1	34.1
5a ($RCI_{JT}a$)	11/16	19.0	43.6
5b ($RCI_{JT}b$)	2/16	49.2	15.1
5c ($RCI_{JT}c$)	11/16	19.6	41.9
6a ($RCI_{CHEL}a$)	15/16	8.4	54.7
6b ($RCI_{CHEL}b$)	4/16	32.4	26.3
6c ($RCI_{CHEL}c$)	16/16	7.8	57.0
7 (RCI_{HSU})	11/16	9.5	58.1
8 (RC_{ID})	2/16	8.9	49.2
9 (RC_{INDIV})	2/16	7.3	44.1
10 (SRB_{MCS})	14/16	10.0	53.1
11 (SRB_{MULT})	13/16	9.5	52.0
12 (SRB_{CH})	14/16	10.0	53.6

Note. Classification accuracy in this table refers to the number of measures (out of 16) that conform to the expected distribution of change classification.

Table 11

Magnitude of prediction intervals for individuals with NCI over the longer test-retest interval

Test	Equations									
	2	3&4	5a&6a	5b&6b	5c&6c	7	8	9	10	11
WMS Information	±0.85	±1.14	±1.61	±2.13	±1.74	±1.31	±1.87	±0.94	±1.65	±1.61
BCR Trial 1	±1.51	±1.87	±2.64	±2.55	±2.81	±2.24	±2.73	±0.80	±2.51	±2.44
BCR Retrieval	±3.69	±3.89	±5.50	±3.72	±6.00	±4.91	±3.30	±3.00	±5.71	±5.43
BCR Acquisition	±0.83	±1.34	±1.90	±2.18	±1.98	±1.36	±1.38	±0.63	±0.50	±0.49
BCR Delay recall	±1.38	±1.61	±2.28	±2.40	±2.44	±1.97	±2.44	±1.16	±1.95	±1.89
RAVLT Trial 1	±1.74	±2.36	±3.34	±2.73	±3.21	±2.72	±2.90	-	±2.50	±2.45
RAVLT Total	±9.47	±9.71	±13.7	±5.64	±13.6	±12.3	±3.92	±5.77	±12.2	±11.9
RAVLT Delay	±3.04	±3.95	±5.58	±3.70	±5.92	±4.63	±2.77	±2.53	±5.17	±5.06
BVRT	±2.34	±2.83	±4.00	±2.97	±3.80	±3.42	±2.90	±1.38	±3.06	±3.00
Similarities	±4.23	±3.53	±4.99	±3.38	±4.92	±4.65	±2.91	±2.46	±4.36	±4.31
Comprehension	±3.03	±3.02	±4.27	±3.17	±4.32	±3.86	±3.20	-	±3.96	±3.91
Token Test	±5.66	±6.79	±9.60	±4.66	±7.69	±8.23	±3.65	±4.54	±6.23	±5.84
FAS	±12.50	±9.36	±13.2	±5.52	±13.1	±12.5	±4.98	±6.35	±11.8	±11.7
Animals	±4.38	±4.59	±6.50	±3.88	±6.48	±5.80	-	±2.55	±5.78	±5.72
Block Design	±4.63	±5.15	±7.29	±3.98	±6.85	±6.40	±4.04	±2.30	±5.55	±5.39
Digit Symbol	±10.51	±6.62	±9.36	±4.78	±9.95	±9.01	-	±3.95	±9.93	±9.25
Average ranking	2	3	9	5	10	7	4	1	7	6

Note. NCI = no cognitive impairment. Prediction intervals are ranked from smallest (1) to largest (10). It was not possible to calculate fixed prediction intervals for equation 12 (SRB_{CH}) as each prediction interval using this method is variable and dependent on the individual's pretest scores. Prediction intervals could not be computed for Animals and WAIS-R Digit Symbol subtests using Equation 8 (RC_{ID}) based on inability to calculate Guttman's reliability coefficients for these measures. Prediction intervals could not be calculated for WAIS-R Comprehension subtest and RAVLT Trial 1 using Equation 9 (RC_{INDIV}) given the negative r_{DD} associated with these measures.

Table 12

Summary of Phase C results

Equation	% of sample with reliable deterioration in memory + other cognitive area	% of sample with no reliably deteriorated scores
2 (SD)	95.0	0.0
3 (RCI)	95.0	0.0
4 (RCI_{SPEER})	95.0	0.0
5a ($RCI_{JT}a$)	90.0	0.0
5b ($RCI_{JT}b$)	100.0	0.0
5c ($RCI_{JT}c$)	90.0	0.0
6a ($RCI_{CHEL}a$)	85.0	0.0
6b ($RCI_{CHEL}b$)	95.0	0.0
6c ($RCI_{CHEL}c$)	80.0	0.0
7 (RCI_{HSU})	90.0	0.0
8 (RC_{ID})	70.0	10.0
9 (RC_{INDIV})	60.0	0.0
10 (SRB_{MCS})	90.0	0.0
11 (SRB_{MULT})	90.0	0.0
12 (SRB_{CH})	90.0	0.0

Table 13

Summary of Phase D results

Equation	Correlation between sum of reliably deteriorated scores and consensus diagnosis	Area under ROC Curve (95% CI interval)
2 (SD)	- 0.460***	0.925 (0.86 – 0.98)***
3 (RCI)	- 0.478***	0.941 (0.88 – 0.99)***
4 (RCI_{SPEER})	- 0.512***	0.967 (0.93 – 1.00)***
5a ($RCI_{JT}a$)	- 0.509***	0.959 (0.92 – 0.99)***
5b ($RCI_{JT}b$)	- 0.465***	0.929 (0.85 – 0.99)***
5c ($RCI_{JT}c$)	- 0.508***	0.958 (0.92 – 0.99)***
6a ($RCI_{CHEL}a$)	- 0.539***	0.972 (0.94 – 0.99)***
6b ($RCI_{CHEL}b$)	- 0.483***	0.943 (0.89 – 0.99)***
6c ($RCI_{CHEL}c$)	- 0.533***	0.963 (0.92 – 1.00)***
7 (RCI_{HSU})	- 0.545***	0.973 (0.94 – 1.00)***
8 (RC_{ID})	- 0.437***	0.887 (0.78 – 0.99)***
9 (RC_{INDIV})	- 0.443***	0.897 (0.82 – 0.97)***
10 (SRB_{MCS})	- 0.534***	0.970 (0.93 – 1.00)***
11 (SRB_{MULT})	- 0.531***	0.967 (0.92 – 1.00)***
12 (SRB_{CH})	- 0.536***	0.971 (0.93 – 1.00)***

Note. ROC = receiver operating characteristics.

* $p < .05$

** $p < .01$

*** $p < .001$

Table 14

Cut-off values for various combinations of sensitivity and specificity using the CSHA neuropsychological battery

Equation	≥ 95% SENS		≥ 95% SPEC		Optimal		
	Cut-off	SPEC	Cut-off	SENS	Cut-off	SENS	SPEC
2 (SD)	≥ 4	69%	≥ 8	70%	≥ 4	95%	69%
3 (RCI)	≥ 4	78%	≥ 7	75%	≥ 5	90%	86%
4 (RCI_{SPEER})	≥ 3	77%	≥ 6	90%	≥ 6	90%	95%
5a ($RCI_{JT}a$)	≥ 2	69%	≥ 5	75%	≥ 3	90%	83%
5b ($RCI_{JT}b$)	≥ 3	51%	≥ 7	75%	≥ 5	90%	80%
5c ($RCI_{JT}c$)	≥ 2	69%	≥ 5	75%	≥ 3	90%	83%
6a ($RCI_{CHEL}a$)	≥ 2	81%	≥ 4	80%	≥ 2	100%	81%
6b ($RCI_{CHEL}b$)	≥ 3	71%	≥ 6	75%	≥ 3	95%	71%
6c ($RCI_{CHEL}c$)	≥ 2	84%	≥ 4	80%	≥ 2	95%	84%
7 (RCI_{HSU})	≥ 3	90%	≥ 5	85%	≥ 4	90%	95%
8 (RC_{ID})	≥ 1	0%	≥ 4	60%	≥ 2	90%	74%
9 (RC_{INDIV})	≥ 1	40%	≥ 5	40%	≥ 2	100%	40%
10 (SRB_{MCS})	≥ 3	89%	≥ 5	90%	≥ 5	90%	96%
11 (SRB_{MULT})	≥ 2	82%	≥ 4	85%	≥ 3	90%	90%
12 (SRB_{CH})	≥ 3	89%	≥ 5	90%	≥ 5	90%	96%

Note. SENS = sensitivity. SPEC = specificity. Optimal cut-off values were those associated with at least 90% sensitivity rates and the highest corresponding specificity value.

Table 15

Summary of Phase E results

Equation	Correlation between sum of reliably deteriorated scores and self-reported loss	Area under ROC Curve (95% CI interval)
2 (SD)	- 0.160*	0.654 (0.54 – 0.75)*
3 (RCI)	- 0.152*	0.646 (0.54 – 0.75)*
4 (RCI_{SPEER})	- 0.182**	0.674 (0.56 – 0.78)**
5a ($RCI_{JT}a$)	- 0.140*	0.633 (0.53 – 0.73)*
5b ($RCI_{JT}b$)	- 0.165**	0.658 (0.55 – 0.76)**
5c ($RCI_{JT}c$)	- 0.132*	0.625 (0.52 – 0.72)*
6a ($RCI_{CHEL}a$)	- 0.130*	0.621 (0.51 – 0.72)*
6b ($RCI_{CHEL}b$)	- 0.124*	0.619 (0.50 – 0.73)
6c ($RCI_{CHEL}c$)	- 0.126*	0.616 (0.51 – 0.72)
7 (RCI_{HSU})	- 0.173**	0.661 (0.54 – 0.77)**
8 (RC_{ID})	- 0.115	0.607 (0.49 – 0.71)
9 (RC_{INDIV})	- 0.189**	0.677 (0.57 – 0.77)**
10 (SRB_{MCS})	- 0.176**	0.665 (0.56 – 0.76)**
11 (SRB_{MULT})	- 0.169**	0.658 (0.56 – 0.75)**
12 (SRB_{CH})	- 0.178**	0.667 (0.56 – 0.76)**

Note. ROC = receiver operating characteristics.

* $p < .05$

** $p < .01$

*** $p < .001$

Table 16

Summary of Phase F results

Equation	Correlation between sum of reliably deteriorated scores and informant-reported loss	Area under ROC Curve (95% CI interval)
2 (SD)	- 0.326***	0.705 (0.63 – 0.77)***
3 (RCI)	- 0.303***	0.691 (0.61 – 0.76)***
4 (RCI_{SPEER})	- 0.322***	0.702 (0.62 – 0.77)***
5a ($RCI_{JT}a$)	- 0.304***	0.689 (0.61 – 0.76)***
5b ($RCI_{JT}b$)	- 0.295***	0.686 (0.61 – 0.76)***
5c ($RCI_{JT}c$)	- 0.302***	0.688 (0.61 – 0.76)***
6a ($RCI_{CHEL}a$)	- 0.311***	0.690 (0.61 – 0.76)***
6b ($RCI_{CHEL}b$)	- 0.285***	0.679 (0.60 – 0.75)***
6c ($RCI_{CHEL}c$)	- 0.298***	0.680 (0.60 – 0.75)***
7 (RCI_{HSU})	- 0.332***	0.703 (0.62 – 0.78)***
8 (RC_{ID})	- 0.229***	0.640 (0.56 – 0.71)***
9 (RC_{INDIV})	- 0.246***	0.651 (0.57 – 0.73)***
10 (SRB_{MCS})	- 0.341***	0.710 (0.63 – 0.78)***
11 (SRB_{MULT})	- 0.327***	0.701 (0.62 – 0.77)***
12 (SRB_{CH})	- 0.346***	0.713 (0.63 – 0.78)***

Note. ROC = receiver operating characteristics.

* $p < .05$

** $p < .01$

*** $p < .001$

Table 17

Summary of Phase G results

Equation	Correlation between sum of reliably deteriorated scores and clinician-rated loss	Area under ROC Curve (95% CI interval)
2 (SD)	- 0.375***	0.755 (0.67 – 0.83)***
3 (RCI)	- 0.357***	0.743 (0.66 – 0.82)***
4 (RCI_{SPEER})	- 0.425***	0.788 (0.71 – 0.86)***
5a ($RCI_{JT}a$)	- 0.384***	0.756 (0.67 – 0.83)***
5b ($RCI_{JT}b$)	- 0.383***	0.761 (0.68 – 0.83)***
5c ($RCI_{JT}c$)	- 0.390***	0.760 (0.68 – 0.83)***
6a ($RCI_{CHEL}a$)	- 0.375***	0.744 (0.66 – 0.82)***
6b ($RCI_{CHEL}b$)	- 0.327***	0.721 (0.63 – 0.80)***
6c ($RCI_{CHEL}c$)	- 0.353***	0.727 (0.64 – 0.81)***
7 (RCI_{HSU})	- 0.415***	0.770 (0.69 – 0.85)***
8 (RC_{ID})	- 0.276***	0.680 (0.59 – 0.76)***
9 (RC_{INDIV})	- 0.364***	0.741 (0.66 – 0.81)***
10 (SRB_{MCS})	- 0.429***	0.781 (0.70 – 0.86)***
11 (SRB_{MULT})	- 0.419***	0.776 (0.69 – 0.85)***
12 (SRB_{CH})	- 0.431***	0.784 (0.70 – 0.86)***

Note. ROC = receiver operating characteristics.

* $p < .05$

** $p < .01$

*** $p < .001$

APPENDIX A

Letter to participants, consent forms, and follow-up letters

LETTER SENT TO POTENTIAL PARTICANTS

Dear «FIRSTNAME» «LASTNAME»,

I am writing to invite you to participate in a special substudy of the Canadian Study of Health and Aging (CSHA-3). In the last few months, you have been kind enough to participate in all of the components of the CSHA including the neuropsychological assessment (this is the component that involved questions about problem solving, language, and memory). The substudy involves repeating the exact same procedure and questions that you completed during the neuropsychological assessment. The reason I am conducting this study is to better understand how much change occurs in cognitively healthy adults, such as yourself, over a relatively brief interval of time (i.e., approximately 1-3 months).

This is an important study because the results will help us to better understand how much change occurs with healthy aging and compare that to the declines that occur with pathological aging processes. The results could allow us to better identify some of the unhealthy changes early in their course and help some seniors get proper treatment as early as possible.

A description of the proposed study is enclosed. A trained interviewer with the project will call you in the next few weeks to arrange a time to come and visit you again. Just as before, the interview will take no more than 1 hour and your answers will be completely confidential.

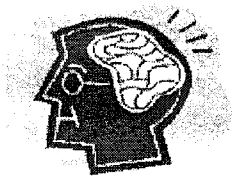
If you have any questions, please call me at (250) 472-4469.

I hope that you will agree to help out.

Sincerely,

Robert Frerichs, Ph.D. candidate
CSHA-3 Project Coordinator

Encl.



PLEASE READ ABOUT THIS OPPORTUNITY FOR YOU!

A neuropsychological assessment, like the one you recently completed as part of the Canadian Study of Health and Aging-3 (CSHA-3), provides a wealth of information about a person's cognitive strengths and weaknesses. However, to detect changes in memory and thinking over time, clinicians need to conduct multiple or repeated assessments.

To distinguish pathological change from the normal change that occurs with healthy aging or day-to-day fluctuations, we need more information from cognitively healthy older adults like you.

Your participation may help clinicians to develop better ways to detect early and potentially important cognitive changes that occur in some older adults.

If you are willing to participate in this study, I will ask you to repeat the same problem solving tasks that you recently completed in the CSHA-3.

Participation may take place in your home and will require no more than one hour of your time.

If you are interested in this study on cognitive changes over time, please call for more information.

Thank you for your interest!

Robert Frerichs
Department of Psychology
University of Victoria

Telephone: (250) 472-4469

CONSENT TO PARTICIPATE IN A STUDY OF COGNITIVE CHANGE

You are being invited to participate in a study entitled "Good? Better? Best? Evaluating methods for detecting cognitive change in older adults" that is being conducted by Mr. Robert Frerichs. Robert Frerichs is a doctoral student in the Department of Psychology at the University of Victoria. You may contact him by calling (250) 472-4469 if you have any questions about the study. As a graduate student, this research is part of the requirements for a Ph.D. degree and it is being conducted under the supervision of Holly Tuokko, Ph.D. She may be reached at (250) 721-6576.

The purpose of this research project is to follow-up participants of the Canadian Study of Health and Aging (CSHA) to determine whether or not their performance changes on select measures over a relatively brief interval (about one month). Research of this type is important because it will advance our understanding of "normal" cognitive changes seen in cognitively healthy older adults. Only by examining "normal" change can we enhance our understanding of the abnormal cognitive changes that occur in some older adults.

You are being asked to participate in this study because you were kind enough to complete the neuropsychological component of the CSHA-3. If you agree to voluntarily participate in this research, your participation will simply include repeating this component of the CSHA-3 study.

As some people may find these tasks tiring please feel free to request breaks as needed. The time required for your participation will be limited to one hour in length.

Your participation in this research must be completely voluntary. If you do decide to participate, you may withdraw at any time without any consequences or any explanation. If you withdraw from the study, you may choose whether or not you want any information gathered to be included in the study.

To protect your anonymity, your information will only be identified using unique code numbers rather than names or other information. Identifying information will be stored separately from all other collected data. No one will ever be identified individually and the results of the study will only be released as a summary along with the information from all other participants. By securely storing the information you provide and limiting access to the principal investigator, your confidentiality and the confidentiality of the data will be protected. Once all required contact with a participant has been concluded, identifying information will be shredded. Anonymous data will be kept for ten years.

It is anticipated that the results of this study will be shared with others to fulfill the requirements of a Ph.D. dissertation. The results of the study may be summarized in a scientific journal or presented at a scholarly meeting.

In addition to being able to contact the researcher and the research supervisor at the above phone numbers, you may verify the ethical approval of this study, or raise any concerns you might have, by contacting the Associate Vice President Research at the University of Victoria (250) 721-7968.

Your signature below indicates that you understand the above conditions of participation in this study and that you have had the opportunity to have your questions answered by the researchers.

PARTICIPANT SIGNATURE _____ DATE _____

A COPY OF THIS CONSENT WILL BE LEFT WITH YOU, AND A COPY WILL BE TAKEN BY THE RESEARCHER

LETTER SENT TO PARTICANTS AFTER STUDY COMPLETION

Dear «FIRSTNAME» «LASTNAME»,

I am writing to express my appreciation for your participation in the Canadian Study of Health and Aging and my dissertation project on normal aging changes. Your time and effort will help make the studies a success. Moreover, the information you provided may help guide future health care decisions and ultimately improve health care for seniors across the country. Thank you very much.

I wish you all the best.

Sincerely,

Robert Frerichs, Ph.D. candidate
CSHA-3 Project Coordinator

APPENDIX B

Coefficients and equations used in calculation of reliable change

Guttman's reliability coefficients and reliability of the difference scores based on neuropsychological test data for persons with No Cognitive Impairment at both CSHA-1 and CSHA-2

Measures	Guttman's reliability coefficients		r_{DD} values	
	CSHA-1	CSHA-2	RC_{ID}	RC_{INDIV}
Memory				
WMS Information	0.52	0.66	0.43	0.67
BCR Trial 1 FR	0.20	0.30	-0.31	0.09
BCR Retrieval	0.76	0.73	0.39	0.50
BCR Acquisition	0.88	0.81	0.84	0.89
BCR Delayed FR	0.31	0.58	-0.04	0.31
RAVLT Trial 1 FR	0.23	0.03	-0.27	-0.23
RAVLT Total Recall	0.91	0.91	0.76	0.77
RAVLT Delayed FR	0.78	0.83	0.69	0.76
BVRT	0.42	0.59	0.08	0.16
Abstract thinking				
WAIS-R Similarities	0.88	0.84	0.45	0.51
Judgment				
WAIS-R Comprehension	0.62	0.62	-0.04	-0.01
Language				
Token Test	0.81	0.79	0.62	0.62
FAS	0.87	0.85	0.34	0.38
Animals	-	-	-	0.19
Construction/Speed				
WAIS-R Block Design	0.48	0.54	-0.06	0.13
WAIS-R Digit Symbol	-	-	-	0.20

Guttman's reliability coefficients and reliability of the difference scores based on neuropsychological test data for persons with No Cognitive Impairment at both CSHA-3 and follow-up

Measures	Guttman's reliability coefficients		r_{DD} values	
	CSHA-1	CSHA-2	RC_{ID}	RC_{INDIV}
Memory				
WMS Information	0.52	0.44	-0.45	-0.08
BCR Trial 1 FR	0.30	0.44	-0.06	0.13
BCR Retrieval	0.78	0.65	0.34	0.51
BCR Acquisition	-	-	-	-
BCR Delayed FR	0.03	0.52	-0.04	0.22
Abstract thinking				
WAIS-R Similarities	0.76	0.75	0.24	0.39
Language				
FAS	0.90	0.86	0.08	0.12
Animals	-	-	-	0.05
Construction/Speed				
WAIS-R Block Design	0.65	0.58	-0.08	-0.03
WAIS-R Digit Symbol	-	-	-	0.08

Regression equations to predict retest scores in Phases B through G based on initial performance

WMS Information (2) = 1.012 + 0.748 * WMS Information (1)
BCR Trial 1 (2) = 3.793 + 0.486 * BCR Trial 1 (1)
BCR Retrieval (2) = 7.101 + 0.684 * BCR Retrieval (1)
BCR Acquisition (2) = 36.135 - 0.006 * BCR Acquisition (1)
BCR Delay (2) = 5.009 + 0.500 * BCR Delay (1)
RAVLT Trial 1 (2) = 3.104 + 0.294 * RAVLT Trial 1 (1)
RAVLT Total (2) = 13.602 + 0.606 * RAVLT Total (1)
RAVLT Delay (2) = 3.869 + 0.419 * RAVLT Delay (1)
BVRT (2) = 7.153 + 0.409 * BVRT (1)
Similarities (2) = 2.163 + 0.745 * Similarities (1)
Comprehension (2) = 3.428 + 0.648 * Comprehension (1)
Token Test (2) = 15.629 + 0.600 * Token Test (1)
FAS (2) = 6.058 + 0.782 * FAS (1)
Animal (2) = 5.753 + 0.590 * Animal (1)
Block Design (2) = 4.516 + 0.469 * Block Design (1)
Digit Symbol (2) = -2.006 + 0.937 * Digit Symbol (1)

Regression equations to predict retest scores in Phases B through G based on initial performance, age, education, and gender

WMS Information (2) = 2.677 + 0.715 * WMS Information (1) - 0.029 * age + 0.050 * education + 0.145 * sex

BCR Trial 1 (2) = 7.475 + 0.470 * BCR Trial 1 (1) - 0.059 * age + 0.048 * education + 0.301 * sex

BCR Retrieval (2) = 16.065 + 0.664 * BCR Retrieval (1) - 0.154 * age + 0.115 * education + 1.371 * sex

BCR Acquisition (2) = 36.337 - 0.007 * BCR Acquisition (1) - 0.005 * age + 0.009 * education + 0.107 * sex

BCR Delay (2) = 7.765 + 0.496 * BCR Delay (1) - 0.039 * age - 0.036 * education + 0.443 * sex

RAVLT Trial 1 (2) = 6.296 + 0.248 * RAVLT Trial 1 (1) - 0.054 * age + 0.032 * education + 0.528 * sex

RAVLT Total (2) = 35.902 + 0.509 * RAVLT Total (1) - 0.313 * age + 0.291 * education + 1.518 * sex

RAVLT Delay (2) = 7.113 + 0.330 * RAVLT Delay (1) - 0.072 * age + 0.168 * education + 0.688 * sex

BVRT (2) = 10.075 + 0.329 * BVRT (1) - 0.037 * age + 0.106 * education - 0.186 * sex

Similarities (2) = 3.343 + 0.682 * Similarities (1) - 0.033 * age + 0.143 * education + 0.177 * sex

Comprehension (2) = 3.374 + 0.585 * Comprehension (1) - 0.007 * age + 0.139 * education - 0.225 * sex

Token Test (2) = 19.374 + 0.538 * Token Test (1) - 0.063 * age + 0.275 * education + 0.385 * sex

FAS (2) = 15.335 + 0.761 * FAS (1) - 0.154 * age + 0.128 * education + 1.140 * sex

Animal (2) = 15.132 + 0.579 * Animal (1) - 0.122 * age + 0.0129 * education + 0.067 * sex

Block Design (2) = 15.299 + 0.389 * Block Design (1) - 0.149 * age + 0.140 * education + 0.026 * sex

Digit Symbol (2) = 29.445 + 0.850 * Digit Symbol (1) - 0.415 * age - 0.005 * education + 2.212 * sex

Equations to predict retest scores in Phase A based on initial performance

WMS Information (2) = 2.021 + 0.596 * WMS Information (1)
BCR Trial 1 (2) = 4.937 + 0.476 * BCR Trial 1 (1)
BCR Retrieval (2) = 13.002 + 0.576 * BCR Retrieval (1)
BCR Delay (2) = 6.023 + 0.449 * BCR Delay (1)
Similarities (2) = 1.397 + 0.844 * Similarities (1)
FAS (2) = 9.899 + 0.747 * FAS (1)
Animal (2) = 4.982 + 0.722 * Animal (1)
Block Design (2) = 4.961 + 0.643 * Block Design (1)
Digit Symbol (2) = 2.392 + 0.946 * Digit Symbol (1)

Equations to predict retest scores in Phase A based on initial performance, age, education, and gender

WMS Information (2) = 5.852 + 0.459 * WMS Information (1) - 0.051 * age + 0.053 * education - 0.063 * sex

BCR Trial 1 (2) = 4.854 + 0.487 * BCR Trial 1 (1) - 0.018 * age + 0.065 * education + 0.375 * sex

BCR Retrieval (2) = 10.596 + 0.574 * BCR Retrieval (1) + 0.022 * age + 0.039 * education + 0.200 * sex

BCR Delay (2) = 7.288 + 0.418 * BCR Delay (1) - 0.034 * age + 0.039 * education + 0.783 * sex

Similarities (2) = -0.554 + 0.778 * Similarities (1) + 0.025 * age + 0.077 * education - 0.132 * sex

FAS (2) = 9.085 + 0.737 * FAS (1) + 0.015 * age + 0.076 * education - 0.669 * sex

Animal (2) = 16.640 + 0.686 * Animal (1) - 0.166 * age + 0.020 * education + 0.655 * sex

Block Design (2) = 20.971 + 0.570 * Block Design (1) - 0.226 * age + 0.142 * education - 0.286 * sex

Digit Symbol (2) = 20.592 + 0.848 * Digit Symbol (1) - 0.247 * age + 0.068 * education + 1.720 * sex

APPENDIX C

Classification data from Phases A, B, and C

Phase A: Wechsler Memory Scale - Information subtest (n = 30)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	6.7	93.3	0.0	-
3 (RCI)	6.7	93.3	0.0	-
4 (RCI_{SPEER})	6.7	86.7	6.7	n.s.
5a ($RCI_{JT}a$)	6.7	93.3	0.0	-
5b ($RCI_{JT}b$)	6.7	93.3	0.0	-
5c ($RCI_{JT}c$)	6.7	93.3	0.0	-
6a ($RCI_{CHEL}a$)	6.7	93.3	0.0	-
6b ($RCI_{CHEL}b$)	3.3	96.7	0.0	-
6c ($RCI_{CHEL}c$)	6.7	93.3	0.0	-
7 (RCI_{HSU})	6.7	90.0	3.3	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	6.7	86.7	6.7	n.s.
11 (SRB_{MULT})	6.7	90.0	3.3	n.s.
12 (SRB_{CH})	6.7	86.7	3.3	n.s.

Phase A: Buschke Cued Recall – Free Recall Trial 1 (n = 30)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	36.7	56.7	6.7	0.001
3 (RCI)	36.7	56.7	6.7	0.001
4 (RCI_{SPEER})	30.0	63.3	6.7	0.001
5a ($RCI_{JT}a$)	10.0	86.7	3.3	n.s.
5b ($RCI_{JT}b$)	10.0	86.7	3.3	n.s.
5c ($RCI_{JT}c$)	10.0	86.7	3.3	n.s.
6a ($RCI_{CHEL}a$)	0.0	93.3	6.7	-
6b ($RCI_{CHEL}b$)	0.0	93.3	6.7	-
6c ($RCI_{CHEL}c$)	0.0	93.3	6.7	-
7 (RCI_{HSU})	0.0	90.0	10.0	-
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	93.3	6.7	-
11 (SRB_{MULT})	0.0	90.0	10.0	-
12 (SRB_{CH})	0.0	93.3	6.7	-

Phase A: Buschke Retrieval (n = 30)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	20.0	76.7	3.3	0.01
3 (RCI)	20.0	76.7	3.3	0.01
4 (RCI_{SPEER})	23.3	73.3	3.3	0.001
5a ($RCI_{JT}a$)	13.3	83.3	3.3	n.s.
5b ($RCI_{JT}b$)	20.0	76.7	3.3	0.01
5c ($RCI_{JT}c$)	13.3	83.3	3.3	n.s.
6a ($RCI_{CHEL}a$)	13.3	83.3	3.3	n.s.
6b ($RCI_{CHEL}b$)	13.3	76.7	10.0	0.05
6c ($RCI_{CHEL}c$)	13.3	83.3	3.3	n.s.
7 (RCI_{HSU})	3.3	86.7	10.0	n.s.
8 (RC_{ID})	13.3	86.7	0.0	-
9 (RC_{INDIV})	20.0	80.0	0.0	-
10 (SRB_{MCS})	3.3	86.7	10.0	n.s.
11 (SRB_{MULT})	3.3	90.0	6.7	n.s.
12 (SRB_{CH})	3.3	90.0	6.7	n.s.

Phase A: Buschke Acquisition (n = 30)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	13.3	86.7	0.0	-
3 (RCI)	13.3	86.7	0.0	-
4 (RCI_{SPEER})	-	-	-	-
5a ($RCI_{JT}a$)	0.0	100.0	0.0	-
5b ($RCI_{JT}b$)	0.0	100.0	0.0	-
5c ($RCI_{JT}c$)	13.3	86.7	0.0	-
6a ($RCI_{CHEL}a$)	-	-	-	-
6b ($RCI_{CHEL}b$)	-	-	-	-
6c ($RCI_{CHEL}c$)	13.3	86.7	0.0	-
7 (RCI_{HSU})	-	-	-	-
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	-	-	-	-
11 (SRB_{MULT})	-	-	-	-
12 (SRB_{CH})	-	-	-	-

Phase A: Buschke Delayed Free Recall (n = 30)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	30.0	63.3	6.7	0.001
3 (RCI)	30.0	63.3	6.7	0.001
4 (RCI_{SPEER})	33.3	60.0	6.7	0.001
5a ($RCI_{JT}a$)	6.7	90.0	3.3	n.s.
5b ($RCI_{JT}b$)	6.7	90.0	3.3	n.s.
5c ($RCI_{JT}c$)	6.7	90.0	3.3	n.s.
6a ($RCI_{CHEL}a$)	6.7	86.7	6.7	n.s.
6b ($RCI_{CHEL}b$)	6.7	86.7	6.7	n.s.
6c ($RCI_{CHEL}c$)	6.7	86.7	6.7	n.s.
7 (RCI_{HSU})	6.7	83.3	10.0	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	93.3	6.7	-
11 (SRB_{MULT})	0.0	93.3	6.7	-
12 (SRB_{CH})	0.0	93.3	6.7	-

Phase A: WAIS-R Similarities subtest (n = 29)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	10.3	82.8	6.9	n.s.
3 (RCI)	10.3	82.8	6.9	n.s.
4 (RCI_{SPEER})	20.7	72.4	6.9	0.001
5a ($RCI_{JT}a$)	0.0	93.1	6.9	-
5b ($RCI_{JT}b$)	10.3	82.8	6.9	n.s.
5c ($RCI_{JT}c$)	0.0	93.1	6.9	-
6a ($RCI_{CHEL}a$)	10.3	82.8	6.9	n.s.
6b ($RCI_{CHEL}b$)	10.3	82.8	6.9	n.s.
6c ($RCI_{CHEL}c$)	0.0	93.1	6.9	-
7 (RCI_{HSU})	3.4	89.7	6.9	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	93.1	6.9	-
11 (SRB_{MULT})	0.0	93.1	6.9	-
12 (SRB_{CH})	0.0	93.1	6.9	-

Phase A: Phonemic fluency - FAS (n = 29)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	0.0	96.6	3.4	-
3 (RCI)	13.8	79.3	6.9	n.s.
4 (RCI_{SPEER})	24.1	72.4	3.4	0.001
5a ($RCI_{JT}a$)	6.9	89.7	3.4	n.s.
5b ($RCI_{JT}b$)	37.9	44.8	17.2	0.001
5c ($RCI_{JT}c$)	6.9	89.7	3.4	n.s.
6a ($RCI_{CHEL}a$)	3.4	93.1	3.4	n.s.
6b ($RCI_{CHEL}b$)	24.1	55.2	20.7	0.001
6c ($RCI_{CHEL}c$)	3.4	93.1	3.4	n.s.
7 (RCI_{HSU})	0.0	96.6	3.4	-
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	96.6	3.4	-
11 (SRB_{MULT})	0.0	96.6	3.4	-
12 (SRB_{CH})	0.0	96.6	3.4	-

Phase A: Category fluency - Animals (n = 29)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	13.8	75.9	10.3	0.05
3 (RCI)	13.8	75.9	10.3	0.05
4 (RCI_{SPEER})	24.1	72.4	3.4	0.001
5a ($RCI_{JT}a$)	13.8	82.8	3.4	n.s.
5b ($RCI_{JT}b$)	13.8	75.9	10.3	0.05
5c ($RCI_{JT}c$)	13.8	82.8	3.4	n.s.
6a ($RCI_{CHEL}a$)	3.4	93.1	3.4	n.s.
6b ($RCI_{CHEL}b$)	13.8	69.0	17.2	0.01
6c ($RCI_{CHEL}c$)	3.4	93.1	3.4	n.s.
7 (RCI_{HSU})	3.4	93.1	3.4	n.s.
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	3.4	96.6	0.0	-
11 (SRB_{MULT})	3.4	96.6	0.0	-
12 (SRB_{CH})	3.4	96.6	0.0	-

Phase A: WAIS-R Block Design subtest (n = 28)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	14.3	78.6	7.1	n.s.
3 (RCI)	14.3	78.6	7.1	n.s.
4 (RCI_{SPEER})	14.3	82.1	3.6	n.s.
5a ($RCI_{JT}a$)	7.1	89.3	3.6	n.s.
5b ($RCI_{JT}b$)	25.0	64.3	10.7	0.001
5c ($RCI_{JT}c$)	7.1	89.3	3.6	n.s.
6a ($RCI_{CHEL}a$)	7.1	85.7	7.1	n.s.
6b ($RCI_{CHEL}b$)	10.7	71.4	17.9	0.01
6c ($RCI_{CHEL}c$)	7.1	85.7	7.1	n.s.
7 (RCI_{HSU})	7.1	85.7	7.1	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	10.7	82.1	7.1	n.s.
11 (SRB_{MULT})	3.6	92.9	3.6	n.s.
12 (SRB_{CH})	10.7	85.7	3.6	n.s.

Phase A: WAIS-R Digit Symbol subtest (n = 27)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	0.0	92.6	7.4	-
3 (RCI)	7.4	85.2	7.4	n.s.
4 (RCI_{SPEER})	14.8	77.8	7.4	n.s.
5a ($RCI_{JT}a$)	0.0	92.6	7.4	-
5b ($RCI_{JT}b$)	14.8	77.8	7.4	n.s.
5c ($RCI_{JT}c$)	0.0	92.6	7.4	-
6a ($RCI_{CHEL}a$)	0.0	92.6	7.4	-
6b ($RCI_{CHEL}b$)	7.4	85.2	7.4	n.s.
6c ($RCI_{CHEL}c$)	0.0	92.6	7.4	-
7 (RCI_{HSU})	3.7	88.9	7.4	n.s.
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	92.6	7.4	-
11 (SRB_{MULT})	0.0	92.6	7.4	-
12 (SRB_{CH})	0.0	92.6	7.4	-

Phase B: Wechsler Memory Scale - Information subtest (n = 166)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	18.1	42.2	39.8	0.001
3 (RCI)	1.8	84.9	13.3	0.001
4 (RCI_{SPEER})	1.8	80.1	18.1	0.001
5a ($RCI_{JT}a$)	1.8	84.9	13.3	0.001
5b ($RCI_{JT}b$)	0.0	96.4	3.6	-
5c ($RCI_{JT}c$)	1.8	84.9	13.3	0.001
6a ($RCI_{CHEL}a$)	1.8	84.9	13.3	0.001
6b ($RCI_{CHEL}b$)	1.8	94.6	3.6	n.s.
6c ($RCI_{CHEL}c$)	1.8	94.6	3.6	n.s.
7 (RCI_{HSU})	1.8	86.7	11.4	0.001
8 (RC_{ID})	0.0	99.4	0.6	-
9 (RC_{INDIV})	1.8	84.9	13.3	0.001
10 (SRB_{MCS})	1.8	91.0	7.2	n.s.
11 (SRB_{MULT})	2.4	89.8	7.8	n.s.
12 (SRB_{CH})	1.8	91.0	7.2	n.s.

Phase B: Buschke Cued Recall – Free Recall Trial 1 (n = 162)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	12.3	62.3	25.3	0.001
3 (RCI)	12.3	62.3	25.3	0.001
4 (RCI_{SPEER})	6.8	78.4	14.8	0.001
5a ($RCI_{JT}a$)	3.1	84.6	12.3	0.001
5b ($RCI_{JT}b$)	3.1	84.6	12.3	0.001
5c ($RCI_{JT}c$)	3.1	84.6	12.3	0.001
6a ($RCI_{CHEL}a$)	3.1	92.0	4.9	n.s.
6b ($RCI_{CHEL}b$)	3.1	92.0	4.9	n.s.
6c ($RCI_{CHEL}c$)	3.1	92.0	4.9	n.s.
7 (RCI_{HSU})	6.2	85.8	8.0	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	4.9	88.3	6.8	n.s.
11 (SRB_{MULT})	4.3	90.7	4.9	n.s.
12 (SRB_{CH})	4.9	88.3	6.8	n.s.

Phase B: Buschke Retrieval (n = 162)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	6.8	69.8	23.5	0.001
3 (RCI)	6.8	69.8	23.5	0.001
4 (RCI_{SPEER})	4.3	77.8	17.9	0.001
5a ($RCI_{JT}a$)	1.9	85.8	12.3	0.001
5b ($RCI_{JT}b$)	6.8	69.8	23.5	0.001
5c ($RCI_{JT}c$)	1.2	89.5	9.3	0.01
6a ($RCI_{CHEL}a$)	6.8	85.2	8.0	n.s.
6b ($RCI_{CHEL}b$)	10.5	77.2	12.3	0.001
6c ($RCI_{CHEL}c$)	3.1	88.9	8.0	n.s.
7 (RCI_{HSU})	4.3	85.2	10.5	0.001
8 (RC_{ID})	0.0	87.7	12.3	-
9 (RC_{INDIV})	0.6	80.2	19.1	0.001
10 (SRB_{MCS})	2.5	90.1	7.4	n.s.
11 (SRB_{MULT})	3.1	88.9	8.0	n.s.
12 (SRB_{CH})	2.5	90.1	7.4	n.s.

Phase B: Buschke Acquisition (n = 162)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	8.0	80.9	11.1	0.001
3 (RCI)	2.5	95.7	1.9	n.s.
4 (RCI_{SPEER})	0.0	97.5	2.5	-
5a ($RCI_{JT}a$)	2.5	95.7	1.9	n.s.
5b ($RCI_{JT}b$)	1.2	96.9	1.9	0.05
5c ($RCI_{JT}c$)	2.5	95.7	1.9	n.s.
6a ($RCI_{CHEL}a$)	2.5	95.7	1.9	n.s.
6b ($RCI_{CHEL}b$)	1.2	96.9	1.9	0.05
6c ($RCI_{CHEL}c$)	2.5	95.7	1.9	n.s.
7 (RCI_{HSU})	0.0	97.5	2.5	-
8 (RC_{ID})	2.5	95.7	1.9	n.s.
9 (RC_{INDIV})	8.0	80.9	11.1	0.001
10 (SRB_{MCS})	0.0	87.7	12.3	-
11 (SRB_{MULT})	0.0	87.7	12.3	-
12 (SRB_{CH})	0.0	87.7	12.3	-

Phase B: Buschke Delayed Free Recall (n = 162)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	8.0	73.5	18.5	0.001
3 (RCI)	8.0	73.5	18.5	0.001
4 (RCI_{SPEER})	11.7	77.2	11.1	0.001
5a ($RCI_{JT}a$)	2.5	90.7	6.8	n.s.
5b ($RCI_{JT}b$)	2.5	90.7	6.8	n.s.
5c ($RCI_{JT}c$)	2.5	90.7	6.8	n.s.
6a ($RCI_{CHEL}a$)	8.0	85.2	6.8	n.s.
6b ($RCI_{CHEL}b$)	2.5	90.7	6.8	n.s.
6c ($RCI_{CHEL}c$)	2.5	90.7	6.8	n.s.
7 (RCI_{HSU})	4.9	87.7	7.4	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	4.9	84.0	11.1	0.01
11 (SRB_{MULT})	3.1	86.4	10.5	0.01
12 (SRB_{CH})	4.3	84.6	11.1	0.01

Phase B: RAVLT Trial 1 (n = 153)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	17.6	59.5	22.9	0.001
3 (RCI)	7.2	83.7	9.2	0.05
4 (RCI_{SPEER})	7.8	86.3	5.9	n.s.
5a ($RCI_{JT}a$)	2.0	93.5	4.6	n.s.
5b ($RCI_{JT}b$)	7.2	83.7	9.2	0.05
5c ($RCI_{JT}c$)	2.0	93.5	4.6	n.s.
6a ($RCI_{CHEL}a$)	2.0	93.5	4.6	n.s.
6b ($RCI_{CHEL}b$)	7.2	83.7	9.2	0.05
6c ($RCI_{CHEL}c$)	2.0	93.5	4.6	n.s.
7 (RCI_{HSU})	5.2	91.5	3.3	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	5.2	90.2	4.6	n.s.
11 (SRB_{MULT})	5.9	89.5	4.6	n.s.
12 (SRB_{CH})	5.2	90.2	4.6	n.s.

Phase B: RAVLT Total score (n = 146)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	6.8	74.7	18.5	0.001
3 (RCI)	6.8	74.7	18.5	0.001
4 (RCI_{SPEER})	7.5	82.9	9.6	0.05
5a ($RCI_{JT}a$)	4.1	87.7	8.2	n.s.
5b ($RCI_{JT}b$)	18.5	39.7	41.8	0.001
5c ($RCI_{JT}c$)	4.1	87.7	8.2	n.s.
6a ($RCI_{CHEL}a$)	5.5	91.1	3.4	n.s.
6b ($RCI_{CHEL}b$)	24.7	50	25.3	0.001
6c ($RCI_{CHEL}c$)	5.5	91.1	3.4	n.s.
7 (RCI_{HSU})	6.2	91.1	2.7	n.s.
8 (RC_{ID})	18.5	37.7	43.8	0.001
9 (RC_{INDIV})	10.3	53.4	36.3	0.001
10 (SRB_{MCS})	6.2	91.1	2.7	n.s.
11 (SRB_{MULT})	5.5	91.1	3.4	n.s.
12 (SRB_{CH})	6.2	91.1	2.7	n.s.

Phase B: RAVLT Delayed Recall (n = 146)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	9.6	68.5	21.9	0.001
3 (RCI)	9.6	68.5	21.9	0.001
4 (RCI_{SPEER})	8.2	80.8	11.0	0.001
5a ($RCI_{JT}a$)	3.4	89.0	7.5	n.s.
5b ($RCI_{JT}b$)	9.6	68.5	21.9	0.001
5c ($RCI_{JT}c$)	3.4	89.0	7.5	n.s.
6a ($RCI_{CHEL}a$)	6.2	90.4	3.4	n.s.
6b ($RCI_{CHEL}b$)	12.3	73.3	14.4	0.001
6c ($RCI_{CHEL}c$)	6.2	90.4	3.4	n.s.
7 (RCI_{HSU})	7.5	86.3	6.2	n.s.
8 (RC_{ID})	6.2	71.9	21.9	0.001
9 (RC_{INDIV})	9.6	68.5	21.9	0.001
10 (SRB_{MCS})	3.4	91.8	4.8	n.s.
11 (SRB_{MULT})	5.5	90.4	4.1	n.s.
12 (SRB_{CH})	3.4	91.8	4.8	n.s.

Phase B: BVRT Multiple Choice (n = 156)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	10.9	76.9	12.2	0.001
3 (RCI)	10.9	76.9	12.2	0.001
4 (RCI_{SPEER})	11.5	82.7	5.8	0.001
5a ($RCI_{JT}a$)	5.1	89.7	5.1	n.s.
5b ($RCI_{JT}b$)	10.9	76.9	12.2	0.001
5c ($RCI_{JT}c$)	5.1	89.7	5.1	n.s.
6a ($RCI_{CHEL}a$)	5.1	92.9	1.9	n.s.
6b ($RCI_{CHEL}b$)	10.9	84.0	5.1	0.01
6c ($RCI_{CHEL}c$)	5.1	89.7	5.1	n.s.
7 (RCI_{HSU})	0.6	93.6	5.8	0.05
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	3.8	89.7	6.4	n.s.
11 (SRB_{MULT})	3.8	91.0	5.1	n.s.
12 (SRB_{CH})	3.8	89.7	6.4	n.s.

Phase B: WAIS-R Similarities subtest (n = 165)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	6.1	86.7	7.3	n.s.
3 (RCI)	11.5	76.4	12.1	0.001
4 (RCI_{SPEER})	11.5	79.4	9.1	0.001
5a ($RCI_{JT}a$)	6.1	86.7	7.3	n.s.
5b ($RCI_{JT}b$)	11.5	76.4	12.1	0.001
5c ($RCI_{JT}c$)	6.1	86.7	7.3	n.s.
6a ($RCI_{CHEL}a$)	3.6	89.1	7.3	n.s.
6b ($RCI_{CHEL}b$)	11.5	76.4	12.1	0.001
6c ($RCI_{CHEL}c$)	3.6	89.1	7.3	n.s.
7 (RCI_{HSU})	4.2	88.5	7.3	n.s.
8 (RC_{ID})	2.4	96.4	1.2	0.05
9 (RC_{INDIV})	6.1	86.7	7.3	n.s.
10 (SRB_{MCS})	4.2	87.9	7.9	n.s.
11 (SRB_{MULT})	5.5	87.9	6.7	n.s.
12 (SRB_{CH})	4.2	87.9	7.9	n.s.

Phase B: WAIS-R Comprehension subtest (n = 165)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	7.3	84.8	7.9	n.s.
3 (RCI)	7.3	84.8	7.9	n.s.
4 (RCI_{SPEER})	12.1	79.4	8.5	0.001
5a ($RCI_{JT}a$)	5.5	89.7	4.8	n.s.
5b ($RCI_{JT}b$)	7.3	84.8	7.9	n.s.
5c ($RCI_{JT}c$)	5.5	89.7	4.8	n.s.
6a ($RCI_{CHEL}a$)	5.5	89.7	4.8	n.s.
6b ($RCI_{CHEL}b$)	7.3	84.8	7.9	n.s.
6c ($RCI_{CHEL}c$)	5.5	89.7	4.8	n.s.
7 (RCI_{HSU})	6.1	88.5	5.5	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	6.1	88.5	5.5	n.s.
11 (SRB_{MULT})	4.2	90.3	5.5	n.s.
12 (SRB_{CH})	6.1	89.7	4.2	n.s.

Phase B: Token Test (n = 158)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	7.6	78.5	13.9	0.001
3 (RCI)	5.1	87.3	7.6	n.s.
4 (RCI_{SPEER})	1.3	91.1	7.6	0.05
5a ($RCI_{JT}a$)	3.2	92.4	4.4	n.s.
5b ($RCI_{JT}b$)	7.6	78.5	13.9	0.001
5c ($RCI_{JT}c$)	5.1	87.3	7.6	n.s.
6a ($RCI_{CHEL}a$)	3.2	94.9	1.9	n.s.
6b ($RCI_{CHEL}b$)	7.6	78.5	13.9	0.001
6c ($RCI_{CHEL}c$)	5.1	90.5	4.4	n.s.
7 (RCI_{HSU})	0.6	94.9	4.4	0.05
8 (RC_{ID})	5.1	81.0	13.9	0.001
9 (RC_{INDIV})	5.1	87.3	7.6	n.s.
10 (SRB_{MCS})	2.5	89.9	7.6	n.s.
11 (SRB_{MULT})	3.8	86.7	9.5	0.05
12 (SRB_{CH})	2.5	89.9	7.6	n.s.

Phase B: Phonemic fluency - FAS (n = 161)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	3.1	91.9	5.0	n.s.
3 (RCI)	9.9	75.8	14.3	0.001
4 (RCI_{SPEER})	11.2	80.1	8.7	0.001
5a ($RCI_{JT}a$)	3.1	93.2	3.7	n.s.
5b ($RCI_{JT}b$)	19.9	53.4	26.7	0.001
5c ($RCI_{JT}c$)	3.1	93.2	3.7	n.s.
6a ($RCI_{CHEL}a$)	3.1	95.0	1.9	n.s.
6b ($RCI_{CHEL}b$)	26.1	50.3	23.6	0.001
6c ($RCI_{CHEL}c$)	3.1	93.2	3.7	n.s.
7 (RCI_{HSU})	4.3	92.5	3.1	n.s.
8 (RC_{ID})	2.5	93.8	3.7	n.s.
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	5.0	90.7	4.3	n.s.
11 (SRB_{MULT})	4.3	91.9	3.7	n.s.
12 (SRB_{CH})	5.0	90.7	4.3	n.s.

Phase B: Category fluency - Animals (n = 163)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	8.6	74.8	16.6	0.001
3 (RCI)	8.6	74.8	16.6	0.001
4 (RCI_{SPEER})	6.1	86.5	7.4	n.s.
5a ($RCI_{JT}a$)	3.7	88.3	8.0	n.s.
5b ($RCI_{JT}b$)	11.0	67.5	21.5	0.001
5c ($RCI_{JT}c$)	3.7	88.3	8.0	n.s.
6a ($RCI_{CHEL}a$)	5.5	89.0	5.5	n.s.
6b ($RCI_{CHEL}b$)	11.0	72.4	16.6	0.001
6c ($RCI_{CHEL}c$)	5.5	89.0	5.5	n.s.
7 (RCI_{HSU})	4.3	92.6	3.1	n.s.
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	4.9	92.0	3.1	n.s.
11 (SRB_{MULT})	4.3	93.9	1.8	n.s.
12 (SRB_{CH})	4.3	92.6	3.1	n.s.

Phase B: WAIS-R Block Design subtest (n = 160)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	4.4	70.6	25.0	0.001
3 (RCI)	3.1	78.8	18.1	0.001
4 (RCI_{SPEER})	2.5	88.1	9.4	0.05
5a ($RCI_{JT}a$)	1.3	88.8	10.0	0.01
5b ($RCI_{JT}b$)	11.9	53.8	34.4	0.001
5c ($RCI_{JT}c$)	1.9	85.6	12.5	0.001
6a ($RCI_{CHEL}a$)	3.1	91.3	5.6	n.s.
6b ($RCI_{CHEL}b$)	16.3	65.6	18.1	0.001
6c ($RCI_{CHEL}c$)	3.1	89.4	7.5	n.s.
7 (RCI_{HSU})	1.9	95.0	3.1	n.s.
8 (RC_{ID})	0.0	100.0	0.0	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	4.4	90.0	5.6	n.s.
11 (SRB_{MULT})	5.0	91.3	3.8	n.s.
12 (SRB_{CH})	4.4	90.6	5.0	n.s.

Phase B: WAIS-R Digit Symbol subtest (n = 153)

Equation	Reliably improved	Not changed	Reliably deteriorated	Chi-square significance
2 (SD)	1.3	87.6	11.1	0.001
3 (RCI)	4.6	61.4	34.0	0.001
4 (RCI_{SPEER})	5.9	66.0	28.1	0.001
5a ($RCI_{JT}a$)	2.6	79.7	17.6	0.001
5b ($RCI_{JT}b$)	7.2	41.2	51.6	0.001
5c ($RCI_{JT}c$)	2.6	79.7	17.6	0.001
6a ($RCI_{CHEL}a$)	5.9	88.2	5.9	n.s.
6b ($RCI_{CHEL}b$)	20.9	54.2	24.8	0.001
6c ($RCI_{CHEL}c$)	5.9	91.5	2.6	n.s.
7 (RCI_{HSU})	6.5	86.9	6.5	n.s.
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	5.2	90.2	4.6	n.s.
11 (SRB_{MULT})	5.2	91.5	3.3	n.s.
12 (SRB_{CH})	4.6	90.8	4.6	n.s.

Phase C: Wechsler Memory Scale - Information subtest (n = 20)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.0	10.0	85.0	n.s.
3 (RCI)	0.0	20.0	80.0	n.s.
4 (RCI_{SPEER})	0.0	15.0	85.0	n.s.
5a ($RCI_{JT}a$)	0.0	20.0	80.0	n.s.
5b ($RCI_{JT}b$)	0.0	45.0	55.0	0.001
5c ($RCI_{JT}c$)	0.0	20.0	80.0	n.s.
6a ($RCI_{CHEL}a$)	0.0	20.0	80.0	n.s.
6b ($RCI_{CHEL}b$)	0.0	45.0	55.0	0.001
6c ($RCI_{CHEL}c$)	0.0	45.0	55.0	0.001
7 (RCI_{HSU})	0.0	15.0	85.0	n.s.
8 (RC_{ID})	0.0	70.0	30.0	0.001
9 (RC_{INDIV})	0.0	20.0	80.0	n.s.
10 (SRB_{MCS})	0.0	20.0	80.0	n.s.
11 (SRB_{MULT})	0.0	30.0	70.0	0.05
12 (SRB_{CH})	0.0	20.0	80.0	n.s.

Phase C: Buschke Cued Recall – Free Recall Trial 1 (n = 20)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.0	15.0	80.0	n.s.
3 (RCI)	5.0	15.0	80.0	n.s.
4 (RCI_{SPEER})	0.0	15.0	85.0	n.s.
5a ($RCI_{JT}a$)	0.0	25.0	75.0	0.05
5b ($RCI_{JT}b$)	0.0	25.0	75.0	0.05
5c ($RCI_{JT}c$)	0.0	25.0	75.0	0.05
6a ($RCI_{CHEL}a$)	0.0	30.0	70.0	0.05
6b ($RCI_{CHEL}b$)	0.0	30.0	70.0	0.05
6c ($RCI_{CHEL}c$)	0.0	30.0	70.0	0.05
7 (RCI_{HSU})	0.0	15.0	85.0	n.s.
8 (RC_{ID})	0.0	100.0	0.0	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	25.0	75.0	0.05
11 (SRB_{MULT})	0.0	20.0	80.0	n.s.
12 (SRB_{CH})	0.0	25.0	75.0	0.05

Phase C: Buschke Retrieval (n = 18)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.6	11.1	83.3	n.s.
3 (RCI)	5.6	11.1	83.3	n.s.
4 (RCI_{SPEER})	0.0	11.1	88.9	n.s.
5a ($RCI_{JT}a$)	0.0	16.7	83.3	n.s.
5b ($RCI_{JT}b$)	5.6	11.1	83.3	n.s.
5c ($RCI_{JT}c$)	0.0	16.7	83.3	n.s.
6a ($RCI_{CHEL}a$)	5.6	16.7	77.8	n.s.
6b ($RCI_{CHEL}b$)	5.6	11.1	83.3	n.s.
6c ($RCI_{CHEL}c$)	5.6	16.7	77.8	n.s.
7 (RCI_{HSU})	0.0	11.1	88.9	n.s.
8 (RC_{ID})	0.0	16.7	83.3	n.s.
9 (RC_{INDIV})	0.0	16.7	83.3	n.s.
10 (SRB_{MCS})	0.0	16.7	83.3	n.s.
11 (SRB_{MULT})	5.6	11.1	83.3	n.s.
12 (SRB_{CH})	0.0	16.7	83.3	n.s.

Phase C: Buschke Acquisition (n = 18)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	16.7	83.3	n.s.
3 (RCI)	0.0	27.8	72.2	0.05
4 (RCI_{SPEER})	0.0	27.8	72.2	0.05
5a ($RCI_{JT}a$)	0.0	27.8	72.2	0.05
5b ($RCI_{JT}b$)	0.0	44.4	55.6	0.001
5c ($RCI_{JT}c$)	0.0	27.8	72.2	0.05
6a ($RCI_{CHEL}a$)	0.0	27.8	72.2	0.05
6b ($RCI_{CHEL}b$)	0.0	44.4	55.6	0.001
6c ($RCI_{CHEL}c$)	0.0	44.4	55.6	0.001
7 (RCI_{HSU})	0.0	27.8	72.2	0.05
8 (RC_{ID})	0.0	27.8	72.2	0.05
9 (RC_{INDIV})	0.0	16.7	83.3	n.s.
10 (SRB_{MCS})	0.0	16.7	83.3	n.s.
11 (SRB_{MULT})	0.0	16.7	83.3	n.s.
12 (SRB_{CH})	0.0	16.7	83.3	n.s.

Phase C: Buschke Delayed Free Recall (n = 18)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	11.1	88.9	n.s.
3 (RCI)	0.0	11.1	88.9	n.s.
4 (RCI_{SPEER})	0.0	11.1	88.9	n.s.
5a ($RCI_{JT}a$)	0.0	27.8	72.2	0.05
5b ($RCI_{JT}b$)	0.0	27.8	72.2	0.05
5c ($RCI_{JT}c$)	0.0	27.8	72.2	0.05
6a ($RCI_{CHEL}a$)	0.0	27.8	72.2	0.05
6b ($RCI_{CHEL}b$)	0.0	27.8	72.2	0.05
6c ($RCI_{CHEL}c$)	0.0	27.8	72.2	0.05
7 (RCI_{HSU})	0.0	16.7	83.3	n.s.
8 (RC_{ID})	0.0	100.0	0.0	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	11.1	88.9	n.s.
11 (SRB_{MULT})	0.0	16.7	83.3	n.s.
12 (SRB_{CH})	0.0	11.1	88.9	n.s.

Phase C: RAVLT Trial 1 (n = 17)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	52.9	47.1	0.001
3 (RCI)	0.0	52.9	47.1	0.001
4 (RCI_{SPEER})	0.0	58.8	41.2	0.001
5a ($RCI_{JT}a$)	0.0	82.4	17.6	0.001
5b ($RCI_{JT}b$)	0.0	52.9	47.1	0.001
5c ($RCI_{JT}c$)	0.0	82.4	17.6	0.001
6a ($RCI_{CHEL}a$)	0.0	82.4	17.6	0.001
6b ($RCI_{CHEL}b$)	0.0	52.9	47.1	0.001
6c ($RCI_{CHEL}c$)	0.0	82.4	17.6	0.001
7 (RCI_{HSU})	0.0	64.7	35.3	0.001
8 (RC_{ID})	0.0	100.0	0.0	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	58.8	41.2	0.001
11 (SRB_{MULT})	0.0	64.7	35.3	0.001
12 (SRB_{CH})	0.0	58.8	41.2	0.001

Phase C: RAVLT Total score (n = 15)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	40.0	60.0	0.01
3 (RCI)	0.0	40.0	60.0	0.01
4 (RCI_{SPEER})	0.0	33.3	66.7	0.05
5a ($RCI_{JT}a$)	0.0	46.7	53.3	0.001
5b ($RCI_{JT}b$)	6.7	20.0	73.3	n.s.
5c ($RCI_{JT}c$)	0.0	46.7	53.3	0.001
6a ($RCI_{CHEL}a$)	0.0	66.7	33.3	0.001
6b ($RCI_{CHEL}b$)	6.7	33.3	60.0	0.01
6c ($RCI_{CHEL}c$)	0.0	66.7	33.3	0.001
7 (RCI_{HSU})	0.0	40.0	60.0	0.01
8 (RC_{ID})	6.7	20.0	73.3	n.s.
9 (RC_{INDIV})	0.0	33.3	66.7	0.05
10 (SRB_{MCS})	0.0	40.0	60.0	0.01
11 (SRB_{MULT})	0.0	40.0	60.0	0.01
12 (SRB_{CH})	0.0	40.0	60.0	0.01

Phase C: RAVLT Delayed Recall (n = 15)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	46.7	53.3	0.001
3 (RCI)	0.0	46.7	53.3	0.001
4 (RCI_{SPEER})	0.0	40.0	60.0	0.01
5a ($RCI_{JT}a$)	0.0	73.3	26.7	0.001
5b ($RCI_{JT}b$)	0.0	46.7	53.3	0.001
5c ($RCI_{JT}c$)	0.0	73.3	26.7	0.001
6a ($RCI_{CHEL}a$)	0.0	80.0	20.0	0.001
6b ($RCI_{CHEL}b$)	0.0	66.7	33.3	0.001
6c ($RCI_{CHEL}c$)	0.0	80.0	20.0	0.001
7 (RCI_{HSU})	0.0	60.0	40.0	0.001
8 (RC_{ID})	0.0	46.7	53.3	0.001
9 (RC_{INDIV})	0.0	46.7	53.3	0.001
10 (SRB_{MCS})	0.0	60.0	40.0	0.001
11 (SRB_{MULT})	0.0	60.0	40.0	0.001
12 (SRB_{CH})	0.0	60.0	40.0	0.001

Phase C: BVRT Multiple Choice (n = 17)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.9	35.3	58.8	0.001
3 (RCI)	5.9	35.3	58.8	0.001
4 (RCI_{SPEER})	5.9	41.2	52.9	0.001
5a ($RCI_{JT}a$)	0.0	41.2	58.8	0.001
5b ($RCI_{JT}b$)	5.9	35.3	58.8	0.001
5c ($RCI_{JT}c$)	0.0	41.2	58.8	0.001
6a ($RCI_{CHEL}a$)	0.0	52.9	47.1	0.001
6b ($RCI_{CHEL}b$)	5.9	35.3	58.8	0.001
6c ($RCI_{CHEL}c$)	0.0	41.2	58.8	0.001
7 (RCI_{HSU})	0.0	47.1	52.9	0.001
8 (RC_{ID})	0.0	100.0	0.0	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	41.2	58.8	0.001
11 (SRB_{MULT})	0.0	41.2	58.8	0.001
12 (SRB_{CH})	0.0	41.2	58.8	0.001

Phase C: WAIS-R Similarities subtest (n = 19)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	73.7	26.3	0.001
3 (RCI)	0.0	57.9	42.1	0.001
4 (RCI_{SPEER})	0.0	57.9	42.1	0.001
5a ($RCI_{JT}a$)	0.0	73.7	26.3	0.001
5b ($RCI_{JT}b$)	0.0	57.9	42.1	0.001
5c ($RCI_{JT}c$)	0.0	73.7	26.3	0.001
6a ($RCI_{CHEL}a$)	0.0	73.7	26.3	0.001
6b ($RCI_{CHEL}b$)	0.0	57.9	42.1	0.001
6c ($RCI_{CHEL}c$)	0.0	73.7	26.3	0.001
7 (RCI_{HSU})	0.0	57.9	42.1	0.001
8 (RC_{ID})	0.0	78.9	21.1	0.001
9 (RC_{INDIV})	0.0	73.7	26.3	0.001
10 (SRB_{MCS})	0.0	57.9	42.1	0.001
11 (SRB_{MULT})	0.0	68.4	31.6	0.001
12 (SRB_{CH})	0.0	57.9	42.1	0.001

Phase C: WAIS-R Comprehension subtest (n = 19)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	52.6	47.4	0.001
3 (RCI)	0.0	52.6	47.4	0.001
4 (RCI_{SPEER})	0.0	42.1	57.9	0.001
5a ($RCI_{JT}a$)	0.0	52.6	47.4	0.001
5b ($RCI_{JT}b$)	0.0	52.6	47.4	0.001
5c ($RCI_{JT}c$)	0.0	52.6	47.4	0.001
6a ($RCI_{CHEL}a$)	0.0	52.6	47.4	0.001
6b ($RCI_{CHEL}b$)	0.0	52.6	47.4	0.001
6c ($RCI_{CHEL}c$)	0.0	52.6	47.4	0.001
7 (RCI_{HSU})	0.0	52.6	47.4	0.001
8 (RC_{ID})	0.0	100.0	0.0	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	52.6	47.4	0.001
11 (SRB_{MULT})	0.0	52.6	47.4	0.001
12 (SRB_{CH})	0.0	52.6	47.4	0.001

Phase C: Token Test (n = 19)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.3	31.6	63.2	0.01
3 (RCI)	0.0	52.6	47.4	0.001
4 (RCI_{SPEER})	0.0	52.6	47.4	0.001
5a ($RCI_{JT}a$)	0.0	63.2	36.8	0.001
5b ($RCI_{JT}b$)	5.3	31.6	63.2	0.01
5c ($RCI_{JT}c$)	0.0	52.6	47.4	0.001
6a ($RCI_{CHEL}a$)	0.0	63.2	36.8	0.001
6b ($RCI_{CHEL}b$)	5.3	31.6	63.2	0.01
6c ($RCI_{CHEL}c$)	0.0	63.2	36.8	0.001
7 (RCI_{HSU})	0.0	52.6	47.4	0.001
8 (RC_{ID})	0.0	36.8	63.2	0.01
9 (RC_{INDIV})	0.0	52.6	47.4	0.001
10 (SRB_{MCS})	0.0	47.4	52.6	0.001
11 (SRB_{MULT})	0.0	47.4	52.6	0.001
12 (SRB_{CH})	0.0	47.4	52.6	0.001

Phase C: Phonemic fluency - FAS (n = 19)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.3	63.2	31.6	0.001
3 (RCI)	5.3	47.4	47.4	0.001
4 (RCI_{SPEER})	0.0	42.1	57.9	0.001
5a ($RCI_{JT}a$)	0.0	68.4	31.6	0.001
5b ($RCI_{JT}b$)	5.3	15.8	78.9	n.s.
5c ($RCI_{JT}c$)	0.0	68.4	31.6	0.001
6a ($RCI_{CHEL}a$)	5.3	68.4	26.3	0.001
6b ($RCI_{CHEL}b$)	5.3	26.3	68.4	0.001
6c ($RCI_{CHEL}c$)	5.3	63.2	31.6	0.001
7 (RCI_{HSU})	0.0	68.4	31.6	0.001
8 (RC_{ID})	0.0	68.4	31.6	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	68.4	31.6	0.001
11 (SRB_{MULT})	0.0	68.4	31.6	0.001
12 (SRB_{CH})	0.0	68.4	31.6	0.001

Phase C: Category fluency - Animals (n = 16)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	56.3	43.8	0.001
3 (RCI)	0.0	56.3	43.8	0.001
4 (RCI_{SPEER})	0.0	43.8	56.3	0.001
5a ($RCI_{JT}a$)	0.0	75.0	25.0	0.001
5b ($RCI_{JT}b$)	0.0	37.5	62.5	0.01
5c ($RCI_{JT}c$)	0.0	75.0	25.0	0.001
6a ($RCI_{CHEL}a$)	0.0	81.3	18.8	0.001
6b ($RCI_{CHEL}b$)	0.0	56.3	43.8	0.001
6c ($RCI_{CHEL}c$)	0.0	81.3	18.8	0.001
7 (RCI_{HSU})	0.0	56.3	43.8	0.001
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	43.8	56.3	0.001
11 (SRB_{MULT})	0.0	62.5	37.5	0.001
12 (SRB_{CH})	0.0	43.8	56.3	0.001

Phase C: WAIS-R Block Design subtest (n = 19)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	5.3	57.9	36.8	0.001
3 (RCI)	5.3	57.9	36.8	0.001
4 (RCI_{SPEER})	5.3	31.6	63.2	0.01
5a ($RCI_{JT}a$)	5.3	57.9	36.8	0.001
5b ($RCI_{JT}b$)	10.5	36.8	52.6	0.001
5c ($RCI_{JT}c$)	5.3	57.9	36.8	0.001
6a ($RCI_{CHEL}a$)	5.3	78.9	15.8	0.001
6b ($RCI_{CHEL}b$)	10.5	52.6	36.8	0.001
6c ($RCI_{CHEL}c$)	5.3	68.4	26.3	0.001
7 (RCI_{HSU})	5.3	68.4	26.3	0.001
8 (RC_{ID})	0	100.0	0.0	0.001
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	5.3	68.4	26.3	0.001
11 (SRB_{MULT})	5.3	68.4	26.3	0.001
12 (SRB_{CH})	5.3	68.4	26.3	0.001

Phase C: WAIS-R Digit Symbol subtest (n = 13)

Equation	Reliably improved	Not changed	Reliably deteriorated	Binomial test significance
2 (SD)	0.0	38.5	61.5	0.01
3 (RCI)	0.0	30.8	69.2	0.05
4 (RCI_{SPEER})	0.0	23.1	76.9	n.s.
5a ($RCI_{JT}a$)	0.0	38.5	61.5	0.01
5b ($RCI_{JT}b$)	0.0	30.8	69.2	0.05
5c ($RCI_{JT}c$)	0.0	38.5	61.5	0.01
6a ($RCI_{CHEL}a$)	0.0	53.8	46.2	0.001
6b ($RCI_{CHEL}b$)	15.4	23.1	61.5	0.01
6c ($RCI_{CHEL}c$)	0.0	53.8	46.2	0.001
7 (RCI_{HSU})	0.0	38.5	61.5	0.01
8 (RC_{ID})	-	-	-	-
9 (RC_{INDIV})	-	-	-	-
10 (SRB_{MCS})	0.0	53.8	46.2	0.001
11 (SRB_{MULT})	0.0	53.8	46.2	0.001
12 (SRB_{CH})	0.0	53.8	46.2	0.001

APPENDIX D

Classification data from Phases D, E, F, and G

Phase D: Wechsler Memory Scale - Information (n = 186; 166 NCI and 20 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	85.0	60.2	20.5	97.1	62.9	8.6	0.282
3 (RCI)	0.001	80.0	86.7	42.1	97.3	86.0	26.2	0.513
4 (RCI_{SPEER})	0.001	85.0	81.9	36.2	97.8	82.3	25.7	0.477
5a ($RCI_{JT}a$)	0.001	80.0	86.7	42.1	97.3	86.0	26.2	0.513
5b ($RCI_{JT}b$)	0.001	55.0	96.4	64.7	94.7	91.9	32.6	0.552
5c ($RCI_{JT}c$)	0.001	80.0	86.7	42.1	97.3	86.0	26.2	0.513
6a ($RCI_{CHEL}a$)	0.001	80.0	86.7	42.1	97.3	86.0	26.2	0.513
6b ($RCI_{CHEL}b$)	0.001	55.0	96.4	64.7	94.7	91.9	32.6	0.552
6c ($RCI_{CHEL}c$)	0.001	55.0	96.4	64.7	94.7	91.9	32.6	0.552
7 (RCI_{HSU})	0.001	85.0	88.6	47.2	98.0	88.2	43.8	0.557
8 (RC_{ID})	0.001	30.0	99.4	85.7	92.2	91.9	70.7	0.479
9 (RC_{INDIV})	0.001	80.0	86.7	42.1	97.3	86.0	26.2	0.513
10 (SRB_{MCS})	0.001	80.0	92.8	57.1	97.5	91.4	51.3	0.630
11 (SRB_{MULT})	0.001	70.0	92.2	51.9	96.2	89.8	27.5	0.547
12 (SRB_{CH})	0.001	80.0	92.8	57.1	97.5	91.4	51.3	0.630

Phase D: Buschke Cued Recall – Free Recall Trial 1 (n = 182; 162 NCI and 20 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	80.0	74.7	28.1	96.8	75.3	11.8	0.369
3 (RCI)	0.001	80.0	74.7	28.1	96.8	75.3	11.8	0.369
4 (RCI_{SPEER})	0.001	85.0	85.2	41.5	97.9	85.2	32.6	0.525
5a ($RCI_{JT}a$)	0.001	75.0	87.7	42.9	96.6	86.3	21.3	0.497
5b ($RCI_{JT}b$)	0.001	75.0	87.7	42.9	96.6	86.3	21.3	0.497
5c ($RCI_{JT}c$)	0.001	75.0	87.7	42.9	96.6	86.3	21.3	0.497
6a ($RCI_{CHEL}a$)	0.001	70.0	95.1	63.6	96.3	92.3	44.9	0.624
6b ($RCI_{CHEL}b$)	0.001	70.0	95.1	63.6	96.3	92.3	44.9	0.624
6c ($RCI_{CHEL}c$)	0.001	70.0	95.1	63.6	96.3	92.3	44.9	0.624
7 (RCI_{HSU})	0.001	85.0	92.0	56.7	98.0	91.2	64.9	0.649
8 (RC_{ID})	-	0.0	100.0	0.0	89.0	89.0	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	75.0	93.2	57.7	96.8	91.2	41.2	0.610
11 (SRB_{MULT})	0.001	80.0	95.1	66.7	97.5	93.4	77.0	0.694
12 (SRB_{CH})	0.001	75.0	93.2	57.7	96.8	91.2	41.2	0.610

Phase D: Buschke Retrieval (n = 180; 162 NCI and 18 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	83.3	76.5	28.3	97.6	77.2	16.3	0.394
3 (RCI)	0.001	83.3	76.5	28.3	97.6	77.2	16.3	0.394
4 (RCI_{SPEER})	0.001	88.9	82.1	35.6	98.5	82.8	36.7	0.492
5a ($RCI_{JT}a$)	0.001	83.3	87.7	42.9	97.9	87.2	35.5	0.538
5b ($RCI_{JT}b$)	0.001	83.3	76.5	28.3	97.6	77.2	16.3	0.394
5c ($RCI_{JT}c$)	0.001	83.3	90.7	50.0	98.0	90.0	49.0	0.596
6a ($RCI_{CHEL}a$)	0.001	77.8	92.0	51.9	97.4	90.6	40.1	0.586
6b ($RCI_{CHEL}b$)	0.001	83.3	87.7	42.9	97.9	87.2	35.5	0.538
6c ($RCI_{CHEL}c$)	0.001	77.8	92.0	51.9	97.4	90.6	40.1	0.586
7 (RCI_{HSU})	0.001	88.9	89.5	48.5	98.6	89.4	68.2	0.608
8 (RC_{ID})	0.001	83.3	87.7	42.9	97.9	87.2	35.5	0.538
9 (RC_{INDIV})	0.001	83.3	80.9	32.6	97.8	81.1	21.1	0.442
10 (SRB_{MCS})	0.001	83.3	92.6	55.6	98.0	91.7	62.5	0.638
11 (SRB_{MULT})	0.001	83.3	92.0	53.6	98.0	91.1	57.3	0.623
12 (SRB_{CH})	0.001	83.3	92.6	55.6	98.0	91.7	62.5	0.638

Phase D: Buschke Acquisition (n = 180; 162 NCI and 18 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	83.3	88.9	45.5	98.0	88.3	40.0	0.560
3 (RCI)	0.001	72.2	98.1	81.3	97.0	95.6	137.8	0.742
4 (RCI_{SPEER})	0.001	72.2	97.5	76.5	96.9	95.0	102.7	0.716
5a ($RCI_{JT}a$)	0.001	72.2	98.1	81.3	97.0	95.6	137.8	0.742
5b ($RCI_{JT}b$)	0.001	55.6	98.1	76.9	95.2	93.9	66.3	0.622
5c ($RCI_{JT}c$)	0.001	72.2	98.1	81.3	97.0	95.6	137.8	0.742
6a ($RCI_{CHEL}a$)	0.001	72.2	98.1	81.3	97.0	95.6	137.8	0.742
6b ($RCI_{CHEL}b$)	0.001	55.6	98.1	76.9	95.2	93.9	66.3	0.622
6c ($RCI_{CHEL}c$)	0.001	55.6	98.1	76.9	95.2	93.9	66.3	0.622
7 (RCI_{HSU})	0.001	72.2	97.5	76.5	96.9	95.0	102.7	0.716
8 (RC_{ID})	0.001	72.2	98.1	81.3	97.0	95.6	137.8	0.742
9 (RC_{INDIV})	0.001	83.3	88.9	45.5	98.0	88.3	40.0	0.560
10 (SRB_{MCS})	0.001	83.3	87.7	42.9	97.9	87.2	35.5	0.538
11 (SRB_{MULT})	0.001	83.3	87.7	42.9	97.9	87.2	35.5	0.538
12 (SRB_{CH})	0.001	83.3	87.7	42.9	97.9	87.2	35.5	0.538

Phase D: Buschke Delayed Free Recall (n = 180; 162 NCI and 18 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	88.9	81.5	34.8	98.5	82.2	35.2	0.484
3 (RCI)	0.001	88.9	81.5	34.8	98.5	82.2	35.2	0.484
4 (RCI_{SPEER})	0.001	88.9	88.9	47.1	98.6	88.9	64.0	0.596
5a ($RCI_{JT}a$)	0.001	72.2	93.2	54.2	96.8	91.1	35.7	0.577
5b ($RCI_{JT}b$)	0.001	72.2	93.2	54.2	96.8	91.1	35.7	0.577
5c ($RCI_{JT}c$)	0.001	72.2	93.2	54.2	96.8	91.1	35.7	0.577
6a ($RCI_{CHEL}a$)	0.001	72.2	93.2	54.2	96.8	91.1	35.7	0.577
6b ($RCI_{CHEL}b$)	0.001	72.2	93.2	54.2	96.8	91.1	35.7	0.577
6c ($RCI_{CHEL}c$)	0.001	72.2	93.2	54.2	96.8	91.1	35.7	0.577
7 (RCI_{HSU})	0.001	83.3	92.6	55.6	98.0	91.7	62.5	0.638
8 (RC_{ID})	-	0.0	100.0	0.0	90.0	90.0	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	88.9	88.9	47.1	98.6	88.9	64.0	0.596
11 (SRB_{MULT})	0.001	83.3	89.5	46.9	98.0	88.9	42.6	0.572
12 (SRB_{CH})	0.001	88.9	88.9	47.1	98.6	88.9	64.0	0.596

Phase D: RAVLT Trial 1 (n = 170; 153 NCI and 17 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	52.9	77.1	20.5	93.7	74.7	3.8	0.206
3 (RCI)	0.001	47.1	90.8	36.4	93.9	86.5	8.8	0.339
4 (RCI_{SPEER})	0.001	41.2	94.1	43.8	93.5	88.8	11.2	0.363
5a ($RCI_{JT}a$)	0.05	17.6	95.4	30.0	91.3	87.6	4.5	0.167
5b ($RCI_{JT}b$)	0.001	47.1	90.8	36.4	93.9	86.5	8.8	0.339
5c ($RCI_{JT}c$)	0.05	17.6	95.4	30.0	91.3	87.6	4.5	0.167
6a ($RCI_{CHEL}a$)	0.05	17.6	95.4	30.0	91.3	87.6	4.5	0.167
6b ($RCI_{CHEL}b$)	0.001	47.1	90.8	36.4	93.9	86.5	8.8	0.339
6c ($RCI_{CHEL}c$)	0.05	17.6	95.4	30.0	91.3	87.6	4.5	0.167
7 (RCI_{HSU})	0.001	35.3	96.7	54.5	93.1	90.6	16.1	0.391
8 (RC_{ID})	-	0.0	100.0	0.0	90.0	90.0	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	41.2	95.4	50.0	93.6	90.0	14.6	0.399
11 (SRB_{MULT})	0.001	35.3	95.4	46.2	93.0	89.4	11.4	0.347
12 (SRB_{CH})	0.001	41.2	95.4	50.0	93.6	90.0	14.6	0.399

Phase D: RAVLT Total score (n = 161; 146 NCI and 15 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	60.0	81.5	25.0	95.2	79.5	6.6	0.290
3 (RCI)	0.001	60.0	81.5	25.0	95.2	79.5	6.6	0.290
4 (RCI_{SPEER})	0.001	66.7	90.4	41.7	96.4	88.2	18.9	0.466
5a ($RCI_{JT}a$)	0.001	53.3	91.8	40.0	95.0	88.2	12.8	0.398
5b ($RCI_{JT}b$)	0.05	73.3	58.2	15.3	95.5	59.6	3.8	0.184
5c ($RCI_{JT}c$)	0.001	53.3	91.8	40.0	95.0	88.2	12.8	0.398
6a ($RCI_{CHEL}a$)	0.001	33.3	96.6	50.0	93.4	90.7	14.1	0.360
6b ($RCI_{CHEL}b$)	0.01	60.0	74.7	19.6	94.8	73.3	4.4	0.223
6c ($RCI_{CHEL}c$)	0.001	33.3	96.6	50.0	93.4	90.7	14.1	0.360
7 (RCI_{HSU})	0.001	60.0	97.3	69.2	95.9	93.8	53.3	0.611
8 (RC_{ID})	0.05	73.3	56.2	14.7	95.3	57.8	3.5	0.172
9 (RC_{INDIV})	0.05	66.7	63.7	15.9	94.9	64.0	3.5	0.181
10 (SRB_{MCS})	0.001	60.0	97.3	69.2	95.9	93.8	53.3	0.611
11 (SRB_{MULT})	0.001	60.0	96.6	64.3	95.9	93.2	42.3	0.584
12 (SRB_{CH})	0.001	60.0	97.3	69.2	95.9	93.8	53.3	0.611

Phase D: RAVLT Delayed Recall (n = 161; 146 NCI and 15 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	53.3	78.1	20.0	94.2	75.8	4.1	0.211
3 (RCI)	0.01	53.3	78.1	20.0	94.2	75.8	4.1	0.211
4 (RCI_{SPEER})	0.001	60.0	89.0	36.0	95.6	86.3	12.2	0.394
5a ($RCI_{JT}a$)	0.05	26.7	92.5	26.7	92.5	86.3	4.5	0.191
5b ($RCI_{JT}b$)	0.01	53.3	78.1	20.0	94.2	75.8	4.1	0.211
5c ($RCI_{JT}c$)	0.05	26.7	92.5	26.7	92.5	86.3	4.5	0.191
6a ($RCI_{CHEL}a$)	0.01	20.0	96.6	37.5	92.2	89.4	7.1	0.222
6b ($RCI_{CHEL}b$)	n.s.	33.3	85.6	19.2	92.6	80.7	3.0	0.150
6c ($RCI_{CHEL}c$)	0.01	20.0	96.6	37.5	92.2	89.4	7.1	0.222
7 (RCI_{HSU})	0.001	40.0	93.8	40.0	93.8	88.8	10.1	0.338
8 (RC_{ID})	0.01	53.3	78.1	20.0	94.2	75.8	4.1	0.211
9 (RC_{INDIV})	0.01	53.3	78.1	20.0	94.2	75.8	4.1	0.211
10 (SRB_{MCS})	0.001	40.0	95.2	46.2	93.9	90.1	13.2	0.376
11 (SRB_{MULT})	0.001	40.0	95.9	50.0	94.0	90.7	15.6	0.397
12 (SRB_{CH})	0.001	40.0	95.2	46.2	93.9	90.1	13.2	0.376

Phase D: BVRT Multiple Choice (n = 173; 156 NCI and 17 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	58.8	87.8	34.5	95.1	85.0	10.3	0.372
3 (RCI)	0.001	58.8	87.8	34.5	95.1	85.0	10.3	0.372
4 (RCI_{SPEER})	0.001	52.9	94.2	50.0	94.8	90.2	18.4	0.460
5a ($RCI_{JT}a$)	0.001	58.8	94.9	55.6	95.5	91.3	26.4	0.524
5b ($RCI_{JT}b$)	0.001	58.8	87.8	34.5	95.1	85.0	10.3	0.372
5c ($RCI_{JT}c$)	0.001	58.8	94.9	55.6	95.5	91.3	26.4	0.524
6a ($RCI_{CHEL}a$)	0.001	47.1	98.1	72.7	94.4	93.1	45.3	0.551
6b ($RCI_{CHEL}b$)	0.001	58.8	94.9	55.6	95.5	91.3	26.4	0.524
6c ($RCI_{CHEL}c$)	0.001	58.8	94.9	55.6	95.5	91.3	26.4	0.524
7 (RCI_{HSU})	0.001	52.9	94.2	50.0	94.8	90.2	18.4	0.460
8 (RC_{ID})	-	0.0	100.0	0.0	90.2	90.2	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	58.8	93.6	50.0	95.4	90.2	20.9	0.488
11 (SRB_{MULT})	0.001	58.8	94.9	55.6	95.5	91.3	26.4	0.524
12 (SRB_{CH})	0.001	58.8	93.6	50.0	95.4	90.2	20.9	0.488

Phase D: WAIS-R Similarities subtest (n = 184; 165 NCI and 19 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	26.3	92.7	29.4	91.6	85.9	4.6	0.200
3 (RCI)	0.01	42.1	87.9	28.6	92.9	81.2	5.3	0.254
4 (RCI_{SPEER})	0.001	42.1	90.9	34.8	93.2	85.9	7.3	0.304
5a ($RCI_{JT}a$)	0.01	26.3	92.7	29.4	91.6	85.9	4.6	0.200
5b ($RCI_{JT}b$)	0.01	42.1	87.9	28.6	92.9	81.2	5.3	0.254
5c ($RCI_{JT}c$)	0.01	26.3	92.7	29.4	91.6	85.9	4.6	0.200
6a ($RCI_{CHEL}a$)	0.01	26.3	92.7	29.4	91.6	85.9	4.6	0.200
6b ($RCI_{CHEL}b$)	0.01	42.1	87.9	28.6	92.9	81.2	5.3	0.254
6c ($RCI_{CHEL}c$)	0.01	26.3	92.7	29.4	91.6	85.9	4.6	0.200
7 (RCI_{HSU})	0.001	42.1	92.7	40.0	93.3	87.5	9.3	0.341
8 (RC_{ID})	0.001	21.1	98.8	66.7	91.6	90.8	21.7	0.340
9 (RC_{INDIV})	0.01	26.3	92.7	29.4	91.6	85.9	4.6	0.200
10 (SRB_{MCS})	0.001	42.1	92.1	38.1	93.3	87.0	8.5	0.328
11 (SRB_{MULT})	0.001	31.6	93.3	35.3	92.2	87.0	6.5	0.262
12 (SRB_{CH})	0.001	42.1	92.1	38.1	93.3	87.0	8.5	0.328

Phase D: WAIS-R Comprehension subtest (n = 184; 165 NCI and 19 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	47.4	92.1	40.9	93.8	87.5	10.5	0.370
3 (RCI)	0.001	47.4	92.1	40.9	93.8	87.5	10.5	0.370
4 (RCI_{SPEER})	0.001	57.9	91.5	44.0	95.0	88.0	14.8	0.439
5a ($RCI_{JT}a$)	0.001	47.4	95.2	52.9	94.0	90.2	17.7	0.447
5b ($RCI_{JT}b$)	0.001	47.4	92.1	40.9	93.8	87.5	10.5	0.370
5c ($RCI_{JT}c$)	0.001	47.4	95.2	52.9	94.0	90.2	17.7	0.447
6a ($RCI_{CHEL}a$)	0.001	47.4	95.2	52.9	94.0	90.2	17.7	0.447
6b ($RCI_{CHEL}b$)	0.001	47.4	92.1	40.9	93.8	87.5	10.5	0.370
6c ($RCI_{CHEL}c$)	0.001	47.4	95.2	52.9	94.0	90.2	17.7	0.447
7 (RCI_{HSU})	0.001	47.4	94.5	50.0	94.0	89.7	15.6	0.429
8 (RC_{ID})	-	0.0	100.0	0.0	89.7	89.7	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	47.4	94.5	50.0	94.0	89.7	15.6	0.429
11 (SRB_{MULT})	0.001	47.4	94.5	50.0	94.0	89.7	15.6	0.429
12 (SRB_{CH})	0.001	47.4	95.8	56.3	94.0	90.8	20.3	0.466

Phase D: Token Test (n = 177; 158 NCI and 19 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	63.2	86.1	35.3	95.1	83.6	10.6	0.387
3 (RCI)	0.001	47.4	92.4	42.9	93.6	87.6	11.0	0.381
4 (RCI_{SPEER})	0.001	47.4	92.4	42.9	93.6	87.6	11.0	0.381
5a ($RCI_{JT}a$)	0.001	36.8	95.6	50.0	92.6	89.3	12.6	0.372
5b ($RCI_{JT}b$)	0.001	63.2	86.1	35.3	95.1	83.6	10.6	0.387
5c ($RCI_{JT}c$)	0.001	47.4	92.4	42.9	93.6	87.6	11.0	0.381
6a ($RCI_{CHEL}a$)	0.001	36.8	98.1	70.0	92.8	91.5	30.1	0.469
6b ($RCI_{CHEL}b$)	0.001	63.2	86.1	35.3	95.1	83.6	10.6	0.387
6c ($RCI_{CHEL}c$)	0.001	36.8	95.6	50.0	92.6	89.3	12.6	0.372
7 (RCI_{HSU})	0.001	47.4	95.6	56.3	93.8	90.4	19.4	0.464
8 (RC_{ID})	0.001	63.2	86.1	35.3	95.1	83.6	10.6	0.387
9 (RC_{INDIV})	0.001	47.4	92.4	42.9	93.6	87.6	11.0	0.381
10 (SRB_{MCS})	0.001	52.9	92.4	45.5	94.2	88.1	13.5	0.423
11 (SRB_{MULT})	0.001	52.6	90.5	40.0	94.1	86.4	10.6	0.383
12 (SRB_{CH})	0.001	52.6	92.4	45.5	94.2	88.1	13.5	0.423

Phase D: Phonemic (FAS) fluency (n = 180; 161 NCI and 19 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	31.6	95.0	42.9	92.2	88.3	8.8	0.305
3 (RCI)	0.001	47.4	85.7	28.1	93.2	81.7	5.4	0.266
4 (RCI_{SPEER})	0.001	57.9	91.3	44.0	94.8	87.8	14.4	0.437
5a ($RCI_{JT}a$)	0.001	31.6	96.3	50.0	92.3	89.4	11.9	0.343
5b ($RCI_{JT}b$)	0.001	78.9	73.3	25.9	96.7	73.9	10.3	0.343
5c ($RCI_{JT}c$)	0.001	31.6	96.3	50.0	92.3	89.4	11.9	0.343
6a ($RCI_{CHEL}a$)	0.001	26.3	98.1	62.5	91.9	90.6	18.8	0.365
6b ($RCI_{CHEL}b$)	0.001	68.4	76.4	25.5	95.3	75.6	7.0	0.306
6c ($RCI_{CHEL}c$)	0.001	31.6	96.3	50.0	92.3	89.4	11.9	0.343
7 (RCI_{HSU})	0.001	31.6	96.9	54.5	92.3	90.0	14.4	0.365
8 (RC_{ID})	0.001	31.6	96.3	50.0	92.3	89.4	11.9	0343
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	31.6	95.7	46.2	92.2	88.9	10.2	0.323
11 (SRB_{MULT})	0.001	31.6	96.3	50.0	92.3	89.4	11.9	0.343
12 (SRB_{CH})	0.001	31.6	95.7	46.2	92.2	88.9	10.2	0.323

Phase D: Category (Animals) fluency (n = 179; 163 NCI and 16 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	43.8	83.4	20.6	93.8	79.9	3.9	0.198
3 (RCI)	0.01	43.8	83.4	20.6	93.8	79.9	3.9	0.198
4 (RCI_{SPEER})	0.001	56.3	92.6	42.9	95.6	89.4	16.1	0.433
5a ($RCI_{JT}a$)	0.05	25.0	92.0	23.5	92.6	86.0	3.8	0.166
5b ($RCI_{JT}b$)	0.001	62.5	78.5	22.2	95.5	77.1	6.1	0.270
5c ($RCI_{JT}c$)	0.05	25.0	92.0	23.5	92.6	86.0	3.8	0.166
6a ($RCI_{CHEL}a$)	0.05	18.8	94.5	25.0	92.2	87.7	3.9	0.151
6b ($RCI_{CHEL}b$)	0.01	43.8	83.4	20.6	93.8	79.9	3.9	0.198
6c ($RCI_{CHEL}c$)	0.05	18.8	94.5	25.0	92.2	87.7	3.9	0.151
7 (RCI_{HSU})	0.001	43.8	96.9	58.3	94.6	92.2	24.6	0.464
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	43.8	96.9	58.3	94.6	92.2	24.6	0.464
11 (SRB_{MULT})	0.001	37.5	98.2	66.7	94.1	92.7	32.0	0.466
12 (SRB_{CH})	0.001	43.8	96.9	58.3	94.6	92.2	24.6	0.464

Phase D: WAIS-R Block Design subtest (n = 179; 160 NCI and 19 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	36.8	75.0	14.9	90.9	70.9	1.8	0.083
3 (RCI)	n.s.	36.8	81.9	19.4	91.6	77.1	2.6	0.144
4 (RCI_{SPEER})	0.001	63.2	90.6	44.4	95.4	87.7	16.6	0.463
5a ($RCI_{JT}a$)	0.01	36.8	90.0	30.4	92.3	84.4	5.3	0.247
5b ($RCI_{JT}b$)	n.s.	52.6	65.6	15.4	92.1	64.2	2.1	0.117
5c ($RCI_{JT}c$)	0.01	36.8	87.5	25.9	92.1	82.1	4.1	0.210
6a ($RCI_{CHEL}a$)	n.s.	15.8	94.4	25.0	90.4	86.0	3.1	0.125
6b ($RCI_{CHEL}b$)	n.s.	36.8	81.9	19.4	91.6	77.1	2.6	0.144
6c ($RCI_{CHEL}c$)	0.01	26.3	92.5	29.4	91.4	85.5	4.4	0.198
7 (RCI_{HSU})	0.001	26.3	96.9	50.0	91.7	89.4	11.1	0.311
8 (RC_{ID})	-	0.0	100.0	0.0	89.4	89.4	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.01	26.3	94.4	35.7	91.5	87.2	6.0	0.237
11 (SRB_{MULT})	0.001	26.3	96.3	45.5	91.7	88.8	9.2	0.289
12 (SRB_{CH})	0.01	26.3	95.0	38.5	91.6	87.7	6.8	0.253

Phase D: WAIS-R Digit Symbol subtest (n = 166; 153 NCI and 13 DEM)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	61.5	88.9	32.0	96.5	86.7	12.8	0.379
3 (RCI)	0.05	69.2	66.0	14.8	96.2	66.3	4.4	0.196
4 (RCI_{SPEER})	0.001	76.9	71.9	18.9	97.3	72.3	8.5	0.281
5a ($RCI_{JT}a$)	0.001	61.5	82.4	22.9	96.2	80.7	7.5	0.289
5b ($RCI_{JT}b$)	n.s.	69.2	48.4	10.2	94.9	50.0	2.1	0.095
5c ($RCI_{JT}c$)	0.001	61.5	82.4	22.9	96.2	80.7	7.5	0.289
6a ($RCI_{CHEL}a$)	0.001	46.2	94.1	40.0	95.4	90.4	13.7	0.377
6b ($RCI_{CHEL}b$)	0.01	61.5	75.2	17.4	95.8	74.1	4.8	0.220
6c ($RCI_{CHEL}c$)	0.001	46.2	97.4	60.0	95.5	93.4	32.0	0.492
7 (RCI_{HSU})	0.001	61.5	93.5	44.4	96.6	91.0	22.9	0.475
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	46.2	95.4	46.2	95.4	91.6	17.9	0.416
11 (SRB_{MULT})	0.001	46.2	96.7	54.5	95.5	92.8	25.4	0.463
12 (SRB_{CH})	0.001	46.2	95.4	46.2	95.4	91.6	17.9	0.416

Phase E: Wechsler Memory Scale - Information (n = 253; 228 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	60.0	51.8	12.0	92.2	52.6	1.6	0.070
3 (RCI)	n.s.	28.0	74.6	10.8	90.4	70.0	1.1	0.017
4 (RCI_{SPEER})	n.s.	48.0	70.2	15.0	92.5	68.0	2.2	0.117
5a ($RCI_{JT}a$)	n.s.	28.0	74.6	10.8	90.4	70.0	1.1	0.017
5b ($RCI_{JT}b$)	n.s.	16.0	88.6	13.3	90.6	81.4	1.5	0.042
5c ($RCI_{JT}c$)	n.s.	28.0	74.6	10.8	90.4	70.0	1.1	0.017
6a ($RCI_{CHEL}a$)	n.s.	28.0	74.6	10.8	90.4	70.0	1.1	0.017
6b ($RCI_{CHEL}b$)	n.s.	16.0	88.6	13.3	90.6	81.4	1.5	0.042
6c ($RCI_{CHEL}c$)	n.s.	16.0	88.6	13.3	90.6	81.4	1.5	0.042
7 (RCI_{HSU})	n.s.	40.0	77.6	16.4	92.2	73.9	2.3	0.123
8 (RC_{ID})	n.s.	4.0	93.9	6.7	89.9	85.0	0.6	-0.027
9 (RC_{INDIV})	n.s.	28.0	74.6	10.8	90.4	70.0	1.1	0.017
10 (SRB_{MCS})	n.s.	20.0	81.6	10.6	90.3	75.5	1.1	0.012
11 (SRB_{MULT})	n.s.	24.0	82.9	13.3	90.9	77.1	1.5	0.054
12 (SRB_{CH})	n.s.	20.0	81.6	10.6	90.3	75.5	1.1	0.012

Phase E: Buschke Cued Recall – Free Recall Trial 1 (n = 248; 223 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	60.0	69.1	17.9	93.9	68.1	3.3	0.185
3 (RCI)	0.01	60.0	69.1	17.9	93.9	68.1	3.3	0.185
4 (RCI_{SPEER})	0.01	48.0	77.1	19.0	93.0	74.2	3.1	0.174
5a ($RCI_{JT}a$)	n.s.	32.0	78.9	14.5	91.2	74.2	1.8	0.079
5b ($RCI_{JT}b$)	n.s.	32.0	78.9	14.5	91.2	74.2	1.8	0.079
5c ($RCI_{JT}c$)	n.s.	32.0	78.9	14.5	91.2	74.2	1.8	0.079
6a ($RCI_{CHEL}a$)	n.s.	12.0	87.4	9.7	89.9	79.8	1.0	-0.005
6b ($RCI_{CHEL}b$)	n.s.	12.0	87.4	9.7	89.9	79.8	1.0	-0.005
6c ($RCI_{CHEL}c$)	n.s.	12.0	87.4	9.7	89.9	79.8	1.0	-0.005
7 (RCI_{HSU})	0.01	40.0	82.1	20.0	92.4	77.8	3.1	0.166
8 (RC_{ID})	-	0.0	100.0	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	24.0	83.9	14.3	90.8	77.8	1.6	0.063
11 (SRB_{MULT})	n.s.	20.0	85.2	13.2	90.5	78.6	1.4	0.043
12 (SRB_{CH})	n.s.	24.0	83.9	14.3	90.8	77.8	1.6	0.063

Phase E: Buschke Retrieval (n = 246; 221 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	64.0	67.4	18.2	93.7	67.1	3.7	0.198
3 (RCI)	0.01	64.0	67.4	18.2	93.7	67.1	3.7	0.198
4 (RCI_{SPEER})	0.001	68.0	71.9	21.5	95.2	71.5	5.5	0.259
5a ($RCI_{JT}a$)	0.01	48.0	76.9	19.0	92.9	74.0	3.1	0.173
5b ($RCI_{JT}b$)	0.01	64.0	67.4	18.2	93.7	67.1	3.7	0.198
5c ($RCI_{JT}c$)	n.s.	36.0	80.5	17.3	91.8	76.0	2.3	0.122
6a ($RCI_{CHEL}a$)	n.s.	32.0	82.4	17.0	91.5	77.2	2.2	0.110
6b ($RCI_{CHEL}b$)	0.01	48.0	76.9	19.0	92.9	74.0	3.1	0.173
6c ($RCI_{CHEL}c$)	n.s.	32.0	82.4	17.0	91.5	77.2	2.2	0.110
7 (RCI_{HSU})	0.01	48.0	78.7	20.3	93.0	75.6	3.4	0.189
8 (RC_{ID})	0.01	48.0	76.9	19.0	92.9	74.0	3.1	0.173
9 (RC_{INDIV})	0.01	56.0	71.9	18.4	93.5	70.3	3.3	0.183
10 (SRB_{MCS})	n.s.	32.0	81.4	16.3	91.4	76.4	2.1	0.102
11 (SRB_{MULT})	n.s.	32.0	81.0	16.0	91.3	76.0	2.0	0.098
12 (SRB_{CH})	n.s.	32.0	81.4	16.3	91.4	76.4	2.1	0.102

Phase E: Buschke Acquisition (n = 246; 221 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	32.0	81.4	16.3	91.4	76.4	2.1	0.102
3 (RCI)	n.s.	8.0	88.2	7.1	89.4	80.1	0.7	-0.036
4 (RCI_{SPEER})	n.s.	8.0	87.3	6.7	89.4	79.3	0.6	-0.043
5a ($RCI_{JT}a$)	n.s.	8.0	88.2	7.1	89.4	80.1	0.7	-0.036
5b ($RCI_{JT}b$)	n.s.	4.0	91.0	4.8	89.3	82.1	0.4	-0.055
5c ($RCI_{JT}c$)	n.s.	8.0	88.2	7.1	89.4	80.1	0.7	-0.036
6a ($RCI_{CHEL}a$)	n.s.	8.0	88.2	7.1	89.4	80.1	0.7	-0.036
6b ($RCI_{CHEL}b$)	n.s.	4.0	91.0	4.8	89.3	82.1	0.4	-0.055
6c ($RCI_{CHEL}c$)	n.s.	4.0	91.0	4.8	89.3	82.1	0.4	-0.055
7 (RCI_{HSU})	n.s.	8.0	87.3	6.7	89.4	79.3	0.6	-0.043
8 (RC_{ID})	n.s.	8.0	88.2	7.1	89.4	80.1	0.7	-0.036
9 (RC_{INDIV})	n.s.	32.0	81.4	16.3	91.4	76.4	2.1	0.102
10 (SRB_{MCS})	n.s.	36.0	78.3	15.8	91.5	74.0	2.0	0.012
11 (SRB_{MULT})	n.s.	36.0	78.3	15.8	91.5	74.0	2.0	0.012
12 (SRB_{CH})	n.s.	36.0	78.3	15.8	91.5	74.0	2.0	0.012

Phase E: Buschke Delayed Free Recall (n = 246; 221 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	48.0	68.3	14.6	92.1	66.3	2.0	0.105
3 (RCI)	n.s.	48.0	68.3	14.6	92.1	66.3	2.0	0.105
4 (RCI_{SPEER})	0.01	56.0	75.6	20.6	93.8	73.6	3.9	0.213
5a ($RCI_{JT}a$)	0.05	36.0	81.0	17.6	91.8	76.4	2.4	0.127
5b ($RCI_{JT}b$)	0.05	36.0	81.0	17.6	91.8	76.4	2.4	0.127
5c ($RCI_{JT}c$)	0.05	36.0	81.0	17.6	91.8	76.4	2.4	0.127
6a ($RCI_{CHEL}a$)	0.05	36.0	81.0	17.6	91.8	76.4	2.4	0.127
6b ($RCI_{CHEL}b$)	0.05	36.0	81.0	17.6	91.8	76.4	2.4	0.127
6c ($RCI_{CHEL}c$)	0.05	36.0	81.0	17.6	91.8	76.4	2.4	0.127
7 (RCI_{HSU})	0.01	48.0	79.6	21.1	93.1	76.4	3.6	0.198
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.01	56.0	75.6	20.6	93.8	73.6	3.9	0.213
11 (SRB_{MULT})	0.01	52.0	76.9	20.3	93.4	74.4	3.6	0.199
12 (SRB_{CH})	0.01	56.0	75.6	20.6	93.8	73.6	3.9	0.213

Phase E: RAVLT Trial 1 (n = 230; 206 no loss and 24 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	33.3	74.8	13.3	90.6	70.4	1.5	0.056
3 (RCI)	n.s.	20.8	87.4	16.1	90.5	80.4	1.8	0.074
4 (RCI_{SPEER})	0.05	25.0	89.8	22.2	91.1	83.0	2.9	0.141
5a ($RCI_{JT}a$)	n.s.	12.5	94.2	20.0	90.2	85.7	2.3	0.083
5b ($RCI_{JT}b$)	n.s.	20.8	87.4	16.1	90.5	80.4	1.8	0.074
5c ($RCI_{JT}c$)	n.s.	12.5	94.2	20.0	90.2	85.7	2.3	0.083
6a ($RCI_{CHEL}a$)	n.s.	12.5	94.2	20.0	90.2	85.7	2.3	0.083
6b ($RCI_{CHEL}b$)	n.s.	20.8	87.4	16.1	90.5	80.4	1.8	0.074
6c ($RCI_{CHEL}c$)	n.s.	12.5	94.2	20.0	90.2	85.7	2.3	0.083
7 (RCI_{HSU})	n.s.	16.7	92.7	21.1	90.5	84.8	2.5	0.104
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	20.8	90.8	20.8	90.8	83.5	2.6	0.116
11 (SRB_{MULT})	0.05	20.8	92.2	23.8	90.9	84.8	3.1	0.139
12 (SRB_{CH})	n.s.	20.8	90.8	20.8	90.8	83.5	2.6	0.116

Phase E: RAVLT Total score (n = 220; 197 no loss and 23 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	47.8	77.2	19.6	92.7	74.1	3.1	0.175
3 (RCI)	0.01	47.8	77.2	19.6	92.7	74.1	3.1	0.175
4 (RCI_{SPEER})	0.01	43.5	88.2	22.2	92.6	78.2	3.6	0.195
5a ($RCI_{JT}a$)	n.s.	26.1	85.8	17.6	90.9	79.5	2.1	0.101
5b ($RCI_{JT}b$)	0.05	69.6	57.4	16.0	94.2	58.6	3.1	0.165
5c ($RCI_{JT}c$)	n.s.	26.1	85.8	17.6	90.9	79.5	2.1	0.101
6a ($RCI_{CHEL}a$)	n.s.	13.0	90.4	13.6	89.9	82.3	1.4	0.035
6b ($RCI_{CHEL}b$)	0.01	56.5	73.1	19.7	93.5	71.4	3.5	0.198
6c ($RCI_{CHEL}c$)	n.s.	13.0	90.4	13.6	89.9	82.3	1.4	0.035
7 (RCI_{HSU})	0.01	30.4	89.8	25.9	91.7	83.6	3.9	0.189
8 (RC_{ID})	0.05	69.6	55.8	15.5	94.0	57.3	2.9	0.156
9 (RC_{INDIV})	0.01	69.6	62.9	18.0	94.7	63.6	3.9	0.203
10 (SRB_{MCS})	0.01	30.4	89.8	25.9	91.7	83.6	3.9	0.189
11 (SRB_{MULT})	0.05	26.1	89.3	22.2	91.2	82.7	3.0	0.144
12 (SRB_{CH})	0.01	30.4	89.8	25.9	91.7	83.6	3.9	0.189

Phase E: RAVLT Delayed Recall (n = 219; 197 no loss and 22 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	36.4	75.6	14.3	91.4	71.7	1.8	0.083
3 (RCI)	n.s.	36.4	75.6	14.3	91.4	71.7	1.8	0.083
4 (RCI_{SPEER})	n.s.	27.3	85.3	17.1	91.3	79.5	2.2	0.103
5a ($RCI_{JT}a$)	n.s.	13.6	90.4	13.6	90.4	82.6	1.5	0.040
5b ($RCI_{JT}b$)	n.s.	36.4	75.6	14.3	91.4	71.7	1.8	0.083
5c ($RCI_{JT}c$)	n.s.	13.6	90.4	13.6	90.4	82.6	1.5	0.040
6a ($RCI_{CHEL}a$)	n.s.	13.6	94.9	23.1	90.8	87.2	3.0	0.109
6b ($RCI_{CHEL}b$)	n.s.	18.2	82.7	10.5	90.1	76.3	1.1	0.007
6c ($RCI_{CHEL}c$)	n.s.	13.6	94.9	23.1	90.8	87.2	3.0	0.109
7 (RCI_{HSU})	n.s.	22.7	90.4	20.8	90.3	83.6	2.8	0.126
8 (RC_{ID})	n.s.	36.4	75.6	14.3	91.4	71.7	1.8	0.083
9 (RC_{INDIV})	n.s.	36.4	75.6	14.3	91.4	71.7	1.8	0.083
10 (SRB_{MCS})	n.s.	18.2	92.4	21.1	91.0	84.9	2.7	0.113
11 (SRB_{MULT})	n.s.	9.1	92.9	12.5	90.1	84.5	1.3	0.023
12 (SRB_{CH})	n.s.	18.2	92.4	21.1	91.0	84.9	2.7	0.113

Phase E: BVRT Multiple Choice (n = 236; 211 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	32.0	81.0	16.7	91.0	75.8	2.0	0.100
3 (RCI)	n.s.	32.0	81.0	16.7	91.0	75.8	2.0	0.100
4 (RCI_{SPEER})	n.s.	20.0	86.7	15.2	90.1	79.7	1.6	0.060
5a ($RCI_{JT}a$)	n.s.	20.0	89.6	18.5	90.4	82.2	2.1	0.093
5b ($RCI_{JT}b$)	n.s.	32.0	81.0	16.7	91.0	75.8	2.0	0.100
5c ($RCI_{JT}c$)	n.s.	20.0	89.6	18.5	90.4	82.2	2.1	0.093
6a ($RCI_{CHEL}a$)	n.s.	8.0	93.4	12.5	89.5	84.3	1.2	0.017
6b ($RCI_{CHEL}b$)	n.s.	20.0	89.6	18.5	90.4	82.2	2.1	0.093
6c ($RCI_{CHEL}c$)	n.s.	20.0	89.6	18.5	90.4	82.2	2.1	0.093
7 (RCI_{HSU})	n.s.	20.0	86.7	15.2	90.1	79.7	1.6	0.060
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	24.0	83.9	15.0	90.3	77.5	1.6	0.065
11 (SRB_{MULT})	n.s.	24.0	85.8	16.7	90.5	79.2	1.9	0.084
12 (SRB_{CH})	n.s.	24.0	83.9	15.0	90.3	77.5	1.6	0.065

Phase E: WAIS-R Similarities subtest (n = 251; 226 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	4.0	89.4	4.0	89.4	80.9	0.4	-0.066
3 (RCI)	n.s.	8.0	83.2	5.0	89.1	75.7	0.4	-0.072
4 (RCI_{SPEER})	n.s.	4.0	84.5	2.8	88.8	76.5	0.2	-0.098
5a ($RCI_{JT}a$)	n.s.	4.0	89.4	4.0	89.4	80.9	0.4	-0.066
5b ($RCI_{JT}b$)	n.s.	8.0	83.2	5.0	89.1	75.7	0.4	-0.072
5c ($RCI_{JT}c$)	n.s.	4.0	89.4	4.0	89.4	80.9	0.4	-0.066
6a ($RCI_{CHEL}a$)	n.s.	4.0	89.4	4.0	89.4	80.9	0.4	-0.066
6b ($RCI_{CHEL}b$)	n.s.	8.0	83.2	5.0	89.1	75.7	0.4	-0.072
6c ($RCI_{CHEL}c$)	n.s.	4.0	89.4	4.0	89.4	80.9	0.4	-0.066
7 (RCI_{HSU})	n.s.	4.0	86.7	3.2	89.1	78.5	0.3	-0.084
8 (RC_{ID})	n.s.	0.0	100.0	0.0	89.6	85.7	1.1	-0.071
9 (RC_{INDIV})	n.s.	4.0	89.4	4.0	89.4	80.9	0.4	-0.066
10 (SRB_{MCS})	n.s.	4.0	85.4	2.9	88.9	77.3	0.2	-0.093
11 (SRB_{MULT})	n.s.	4.0	88.5	3.7	89.3	80.1	0.3	-0.073
12 (SRB_{CH})	n.s.	4.0	85.4	2.9	88.9	77.3	0.2	-0.093

Phase E: WAIS-R Comprehension subtest (n = 251; 226 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	4.0	85.8	3.0	89.0	77.7	0.3	-0.090
3 (RCI)	n.s.	4.0	85.8	3.0	89.0	77.7	0.3	-0.090
4 (RCI_{SPEER})	n.s.	8.0	82.7	4.9	89.0	75.3	0.4	-0.075
5a ($RCI_{JT}a$)	n.s.	4.0	91.2	4.8	89.6	82.5	0.4	-0.052
5b ($RCI_{JT}b$)	n.s.	4.0	85.8	3.0	89.0	77.7	0.3	-0.090
5c ($RCI_{JT}c$)	n.s.	4.0	91.2	4.8	89.6	82.5	0.4	-0.052
6a ($RCI_{CHEL}a$)	n.s.	4.0	91.2	4.8	89.6	82.5	0.4	-0.052
6b ($RCI_{CHEL}b$)	n.s.	4.0	85.8	3.0	89.0	77.7	0.3	-0.090
6c ($RCI_{CHEL}c$)	n.s.	4.0	91.2	4.8	89.6	82.5	0.4	-0.052
7 (RCI_{HSU})	n.s.	4.0	87.6	3.4	89.2	79.3	0.3	-0.079
8 (RC_{ID})	-	0.0	100.0	0.0	90.0	90.0	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	4.0	88.9	3.8	89.3	80.5	0.3	-0.069
11 (SRB_{MULT})	n.s.	4.0	89.8	4.2	89.4	81.3	0.4	-0.063
12 (SRB_{CH})	n.s.	4.0	89.8	4.2	89.4	81.3	0.4	-0.063

Phase E: Token Test (n = 241; 216 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	8.0	79.6	4.3	88.2	72.2	0.3	-0.096
3 (RCI)	n.s.	8.0	88.0	7.1	89.2	79.7	0.6	-0.038
4 (RCI_{SPEER})	n.s.	12.0	87.0	9.7	89.5	79.3	0.9	-0.009
5a ($RCI_{JT}a$)	n.s.	4.0	92.1	5.6	89.2	83.0	0.5	-0.045
5b ($RCI_{JT}b$)	n.s.	8.0	79.6	4.3	88.2	72.2	0.3	-0.096
5c ($RCI_{JT}c$)	n.s.	8.0	88.0	7.1	89.2	79.7	0.6	-0.038
6a ($RCI_{CHEL}a$)	n.s.	4.0	94.4	7.7	89.5	85.1	0.7	-0.021
6b ($RCI_{CHEL}b$)	n.s.	8.0	79.6	4.3	88.2	72.2	0.3	-0.096
6c ($RCI_{CHEL}c$)	n.s.	4.0	92.1	5.6	89.2	83.0	0.5	-0.045
7 (RCI_{HSU})	n.s.	12.0	90.3	12.5	89.9	82.2	1.3	0.023
8 (RC_{ID})	n.s.	8.0	79.6	4.3	88.2	72.2	0.3	-0.096
9 (RC_{INDIV})	n.s.	8.0	88.0	7.1	89.2	79.7	0.6	-0.038
10 (SRB_{MCS})	n.s.	12.0	86.1	9.1	89.4	78.4	0.8	-0.017
11 (SRB_{MULT})	n.s.	12.0	84.7	8.3	89.3	77.2	0.8	-0.028
12 (SRB_{CH})	n.s.	12.0	86.1	9.1	89.4	78.4	0.8	-0.017

Phase E: Phonemic (FAS) fluency (n = 246; 221 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	8.0	91.4	9.5	89.8	82.9	0.9	-0.006
3 (RCI)	n.s.	16.0	80.1	8.3	89.4	73.6	0.8	-0.030
4 (RCI_{SPEER})	n.s.	16.0	85.5	11.1	90.0	78.5	1.1	0.013
5a ($RCI_{JT}a$)	n.s.	4.0	92.8	5.9	89.5	83.7	0.5	-0.039
5b ($RCI_{JT}b$)	n.s.	40.0	65.6	11.6	90.6	63.0	1.3	0.036
5c ($RCI_{JT}c$)	n.s.	4.0	92.8	5.9	89.5	83.7	0.5	-0.039
6a ($RCI_{CHEL}a$)	n.s.	0.0	95.5	0.0	89.4	85.8	1.0	-0.069
6b ($RCI_{CHEL}b$)	n.s.	36.0	70.1	12.0	90.6	66.7	1.3	0.040
6c ($RCI_{CHEL}c$)	n.s.	4.0	92.8	5.9	89.5	83.7	0.5	-0.039
7 (RCI_{HSU})	n.s.	4.0	92.8	5.9	89.5	83.7	0.5	-0.039
8 (RC_{ID})	n.s.	4.0	92.8	5.9	89.5	83.7	0.5	-0.039
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	4.0	91.9	5.3	89.4	82.9	0.5	-0.047
11 (SRB_{MULT})	n.s.	4.0	92.8	5.9	89.5	83.7	0.5	-0.039
12 (SRB_{CH})	n.s.	4.0	91.9	5.3	89.4	82.9	0.5	-0.047

Phase E: Category (Animals) fluency (n = 246; 221 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	28.0	80.5	14.0	90.8	75.2	1.6	0.064
3 (RCI)	n.s.	28.0	80.5	14.0	90.8	75.2	1.6	0.064
4 (RCI_{SPEER})	n.s.	20.0	84.6	12.8	90.3	78.0	1.4	0.038
5a ($RCI_{JT}a$)	n.s.	8.0	89.1	7.7	89.5	80.9	0.7	-0.028
5b ($RCI_{JT}b$)	n.s.	32.0	73.8	12.1	90.6	69.5	1.3	0.039
5c ($RCI_{JT}c$)	n.s.	8.0	89.1	7.7	89.5	80.9	0.7	-0.028
6a ($RCI_{CHEL}a$)	n.s.	8.0	91.9	10.0	89.8	83.3	1.0	-0.002
6b ($RCI_{CHEL}b$)	n.s.	28.0	80.5	14.0	90.8	75.2	1.6	0.064
6c ($RCI_{CHEL}c$)	n.s.	8.0	91.9	10.0	89.8	83.3	1.0	-0.002
7 (RCI_{HSU})	n.s.	12.0	92.3	15.0	90.3	84.1	1.6	0.048
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	12.0	92.3	15.0	90.3	84.1	1.6	0.048
11 (SRB_{MULT})	n.s.	12.0	94.1	18.8	90.4	85.8	2.2	0.075
12 (SRB_{CH})	n.s.	12.0	92.3	15.0	90.3	84.1	1.6	0.048

Phase E: WAIS-R Block Design subtest (n = 243; 218 no loss and 25 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	28.0	71.6	10.1	89.7	67.1	1.0	-0.003
3 (RCI)	n.s.	16.0	79.4	8.2	89.2	72.8	0.7	-0.035
4 (RCI_{SPEER})	n.s.	16.0	82.6	9.5	89.6	75.7	0.9	-0.011
5a ($RCI_{JT}a$)	n.s.	12.0	87.2	9.7	89.6	79.4	0.9	-0.008
5b ($RCI_{JT}b$)	n.s.	48.0	61.0	12.4	91.1	59.7	1.4	0.056
5c ($RCI_{JT}c$)	n.s.	12.0	84.9	8.3	89.4	77.4	0.7	-0.027
6a ($RCI_{CHEL}a$)	n.s.	4.0	93.6	6.7	89.5	84.4	0.6	-0.031
6b ($RCI_{CHEL}b$)	n.s.	16.0	79.4	8.2	89.2	72.8	0.7	-0.035
6c ($RCI_{CHEL}c$)	n.s.	8.0	90.8	9.1	89.6	82.3	0.8	-0.012
7 (RCI_{HSU})	n.s.	8.0	93.6	12.5	89.9	84.8	1.3	0.019
8 (RC_{ID})	-	0.0	100.0	0.0	89.7	89.7	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	12.0	89.0	11.1	89.8	81.1	1.1	0.010
11 (SRB_{MULT})	n.s.	4.0	90.8	4.8	89.2	81.9	0.4	-0.056
12 (SRB_{CH})	n.s.	12.0	89.4	11.5	89.9	81.5	1.2	0.014

Phase E: WAIS-R Digit Symbol subtest (n = 223; 199 no loss and 24 self-reported loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	16.7	85.4	12.1	89.5	78.0	1.2	0.018
3 (RCI)	n.s.	41.7	64.3	12.3	90.1	61.9	1.3	0.039
4 (RCI_{SPEER})	n.s.	41.7	66.3	13.0	90.4	63.7	1.4	0.052
5a ($RCI_{JT}a$)	n.s.	25.0	80.9	13.6	89.9	74.9	1.4	0.046
5b ($RCI_{JT}b$)	n.s.	54.2	46.7	10.9	89.4	47.5	1.0	0.006
5c ($RCI_{JT}c$)	n.s.	25.0	80.9	13.6	89.9	74.9	1.4	0.046
6a ($RCI_{CHEL}a$)	n.s.	12.5	91.5	15.0	89.7	83.0	1.5	0.043
6b ($RCI_{CHEL}b$)	n.s.	29.2	74.9	12.3	89.8	70.0	1.2	0.029
6c ($RCI_{CHEL}c$)	n.s.	4.2	93.5	7.1	89.0	83.9	0.6	-0.030
7 (RCI_{HSU})	n.s.	16.7	89.9	16.7	89.9	82.1	1.8	0.066
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	12.5	92.5	16.7	89.8	83.9	1.8	0.056
11 (SRB_{MULT})	n.s.	8.3	92.5	11.8	89.3	83.4	1.1	0.009
12 (SRB_{CH})	n.s.	12.5	92.5	16.7	89.8	83.9	1.8	0.056

Phase F: Wechsler Memory Scale - Information (n = 251; 178 no loss and 73 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	68.5	57.3	39.7	81.6	60.6	2.9	0.234
3 (RCI)	0.01	38.4	79.2	43.1	75.8	67.3	2.4	0.182
4 (RCI_{SPEER})	0.01	46.6	74.2	42.5	77.2	66.1	2.5	0.202
5a ($RCI_{JT}a$)	0.01	38.4	79.2	43.1	75.8	67.3	2.4	0.182
5b ($RCI_{JT}b$)	0.01	21.9	92.1	53.3	74.2	71.7	3.3	0.197
5c ($RCI_{JT}c$)	0.01	38.4	79.2	43.1	75.8	67.3	2.4	0.182
6a ($RCI_{CHEL}a$)	0.01	38.4	79.2	43.1	75.8	67.3	2.4	0.182
6b ($RCI_{CHEL}b$)	0.01	21.9	92.1	53.3	74.2	71.7	3.3	0.197
6c ($RCI_{CHEL}c$)	0.01	21.9	92.1	53.3	74.2	71.7	3.3	0.197
7 (RCI_{HSU})	0.001	42.5	82.0	49.2	77.7	70.5	3.4	0.256
8 (RC_{ID})	0.01	12.3	96.6	60.0	72.9	72.1	4.0	0.172
9 (RC_{INDIV})	0.01	38.4	79.2	43.1	75.8	67.3	2.4	0.182
10 (SRB_{MCS})	0.001	34.2	87.1	52.1	76.4	71.7	3.5	0.246
11 (SRB_{MULT})	0.001	31.5	87.6	51.1	75.7	71.3	3.3	0.227
12 (SRB_{CH})	0.001	34.2	87.1	52.1	76.4	71.7	3.5	0.246

Phase F: Buschke Cued Recall – Free Recall Trial 1 (n = 245; 174 no loss and 71 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	59.2	77.0	51.2	82.2	71.8	4.9	0.348
3 (RCI)	0.001	59.2	77.0	51.2	82.2	71.8	4.9	0.348
4 (RCI_{SPEER})	0.001	50.7	85.1	58.1	80.9	75.1	5.9	0.373
5a ($RCI_{JT}a$)	0.001	46.5	88.5	62.3	80.2	76.3	6.7	0.385
5b ($RCI_{JT}b$)	0.001	46.5	88.5	62.3	80.2	76.3	6.7	0.385
5c ($RCI_{JT}c$)	0.001	46.5	88.5	62.3	80.2	76.3	6.7	0.385
6a ($RCI_{CHEL}a$)	0.001	31.0	94.3	68.8	77.0	75.9	7.4	0.340
6b ($RCI_{CHEL}b$)	0.001	31.0	94.3	68.8	77.0	75.9	7.4	0.340
6c ($RCI_{CHEL}c$)	0.001	31.0	94.3	68.8	77.0	75.9	7.4	0.340
7 (RCI_{HSU})	0.001	47.9	90.8	68.0	81.0	78.4	9.0	0.436
8 (RC_{ID})	-	0.0	100.0	0.0	71.0	71.0	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	39.4	92.0	66.7	78.8	76.7	7.4	0.378
11 (SRB_{MULT})	0.001	35.2	92.5	65.8	77.8	75.9	6.7	0.348
12 (SRB_{CH})	0.001	39.4	92.0	66.7	78.8	76.7	7.4	0.378

Phase F: Buschke Retrieval (n = 243; 174 no loss and 69 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	58.0	73.6	46.5	81.5	69.1	3.8	0.297
3 (RCI)	0.001	58.0	73.6	46.5	81.5	69.1	3.8	0.297
4 (RCI_{SPEER})	0.001	60.9	79.9	54.5	83.7	74.5	6.2	0.395
5a ($RCI_{JT}a$)	0.001	47.8	83.3	53.2	80.1	73.3	4.6	0.322
5b ($RCI_{JT}b$)	0.001	58.0	73.6	46.5	81.5	69.1	3.8	0.297
5c ($RCI_{JT}c$)	0.001	42.0	86.8	55.8	79.1	74.1	4.8	0.317
6a ($RCI_{CHEL}a$)	0.001	42.0	89.7	61.7	79.6	76.1	6.3	0.362
6b ($RCI_{CHEL}b$)	0.001	47.8	83.3	53.2	80.1	73.3	4.6	0.322
6c ($RCI_{CHEL}c$)	0.001	42.0	89.7	61.7	79.6	76.1	6.3	0.362
7 (RCI_{HSU})	0.001	49.3	85.6	57.6	81.0	75.3	5.8	0.367
8 (RC_{ID})	0.001	47.8	83.3	53.2	80.1	73.3	4.6	0.322
9 (RC_{INDIV})	0.001	55.1	78.7	50.7	81.5	72.0	4.6	0.330
10 (SRB_{MCS})	0.001	43.5	89.1	61.2	79.9	76.1	6.3	0.366
11 (SRB_{MULT})	0.001	42.0	87.9	58.0	79.3	74.9	5.3	0.334
12 (SRB_{CH})	0.001	43.5	89.1	61.2	79.9	76.1	6.3	0.366

Phase F: Buschke Acquisition (n = 243; 174 no loss and 69 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	40.6	87.9	57.1	78.9	74.5	5.0	0.320
3 (RCI)	0.001	26.1	93.7	62.1	76.2	74.5	5.2	0.275
4 (RCI_{SPEER})	0.001	26.1	92.5	58.1	75.9	73.7	4.4	0.252
5a ($RCI_{JT}a$)	0.001	26.1	93.7	62.1	76.2	74.5	5.2	0.275
5b ($RCI_{JT}b$)	0.001	23.2	96.6	72.7	76.0	75.7	8.5	0.310
5c ($RCI_{JT}c$)	0.001	26.1	93.7	62.1	76.2	74.5	5.2	0.275
6a ($RCI_{CHEL}a$)	0.001	26.1	93.7	62.1	76.2	74.5	5.2	0.275
6b ($RCI_{CHEL}b$)	0.001	23.2	96.6	72.7	76.0	75.7	8.5	0.310
6c ($RCI_{CHEL}c$)	0.001	23.2	96.6	72.7	76.0	75.7	8.5	0.310
7 (RCI_{HSU})	0.001	26.1	92.5	58.1	75.9	73.7	4.4	0.252
8 (RC_{ID})	0.001	26.1	93.7	62.1	76.2	74.5	5.2	0.275
9 (RC_{INDIV})	0.001	40.6	87.9	57.1	78.9	74.5	5.0	0.320
10 (SRB_{MCS})	0.001	43.5	84.5	52.6	79.0	72.8	4.2	0.298
11 (SRB_{MULT})	0.001	43.5	84.5	52.6	79.0	72.8	4.2	0.298
12 (SRB_{CH})	0.001	43.5	84.5	52.6	79.0	72.8	4.2	0.298

Phase F: Buschke Delayed Free Recall (n = 243; 173 no loss and 70 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	55.7	75.7	48.1	80.9	70.0	3.9	0.302
3 (RCI)	0.001	55.7	75.7	48.1	80.9	70.0	3.9	0.302
4 (RCI_{SPEER})	0.001	52.9	82.7	55.2	81.3	74.1	5.3	0.360
5a ($RCI_{JT}a$)	0.001	41.4	87.9	58.0	78.8	74.5	5.1	0.328
5b ($RCI_{JT}b$)	0.001	41.4	87.9	58.0	78.8	74.5	5.1	0.328
5c ($RCI_{JT}c$)	0.001	41.4	87.9	58.0	78.8	74.5	5.1	0.328
6a ($RCI_{CHEL}a$)	0.001	41.4	87.9	58.0	78.8	74.5	5.1	0.328
6b ($RCI_{CHEL}b$)	0.001	41.4	87.9	58.0	78.8	74.5	5.1	0.328
6c ($RCI_{CHEL}c$)	0.001	41.4	87.9	58.0	78.8	74.5	5.1	0.328
7 (RCI_{HSU})	0.001	47.1	86.1	57.9	80.1	74.9	5.5	0.356
8 (RC_{ID})	-	0.0	100.0	0.0	71.2	71.2	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	52.9	82.7	55.2	81.3	74.1	5.3	0.360
11 (SRB_{MULT})	0.001	51.4	84.4	57.1	81.1	74.9	5.7	0.370
12 (SRB_{CH})	0.001	52.9	82.7	55.2	81.3	74.1	5.3	0.360

Phase F: RAVLT Trial 1 (n = 228; 163 no loss and 65 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	27.7	74.2	30.0	72.0	61.0	1.1	0.020
3 (RCI)	n.s.	18.5	88.3	38.7	73.1	68.4	0.8	0.090
4 (RCI_{SPEER})	n.s.	16.9	90.2	40.7	73.1	69.3	1.9	0.099
5a ($RCI_{JT}a$)	n.s.	9.2	95.1	42.9	72.4	70.6	2.0	0.081
5b ($RCI_{JT}b$)	n.s.	18.5	88.3	38.7	73.1	68.4	0.8	0.090
5c ($RCI_{JT}c$)	n.s.	9.2	95.1	42.9	72.4	70.6	2.0	0.081
6a ($RCI_{CHEL}a$)	n.s.	9.2	95.1	42.9	72.4	70.6	2.0	0.081
6b ($RCI_{CHEL}b$)	n.s.	18.5	88.3	38.7	73.1	68.4	0.8	0.090
6c ($RCI_{CHEL}c$)	n.s.	9.2	95.1	42.9	72.4	70.6	2.0	0.081
7 (RCI_{HSU})	0.05	15.4	94.5	52.6	73.7	71.9	3.1	0.161
8 (RC_{ID})	-	0.0	100.0	0.0	71.5	71.5	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.05	16.9	92.0	45.8	73.5	70.6	2.4	0.132
11 (SRB_{MULT})	0.05	15.4	93.3	47.6	73.4	71.1	2.5	0.135
12 (SRB_{CH})	0.05	16.9	92.0	45.8	73.5	70.6	2.4	0.132

Phase F: RAVLT Total score (n = 219; 158 no loss and 61 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.05	36.1	79.1	40.0	76.2	67.1	2.1	0.157
3 (RCI)	0.05	36.1	79.1	40.0	76.2	67.1	2.1	0.157
4 (RCI_{SPEER})	0.05	29.5	83.5	70.9	75.4	68.5	2.1	0.146
5a ($RCI_{JT}a$)	0.05	24.6	88.0	44.1	75.1	70.3	2.4	0.156
5b ($RCI_{JT}b$)	n.s.	50.8	57.0	31.3	75.0	55.3	1.4	0.070
5c ($RCI_{JT}c$)	0.05	24.6	88.0	44.1	75.1	70.3	2.4	0.156
6a ($RCI_{CHEL}a$)	n.s.	16.4	92.4	45.5	74.1	71.2	2.4	0.131
6b ($RCI_{CHEL}b$)	n.s.	37.7	73.4	35.4	75.3	63.5	1.8	0.109
6c ($RCI_{CHEL}c$)	n.s.	16.4	92.4	45.5	74.1	71.2	2.4	0.131
7 (RCI_{HSU})	0.001	27.9	94.3	65.4	77.2	75.8	6.4	0.307
8 (RC_{ID})	n.s.	52.5	55.7	31.4	75.2	54.8	1.4	0.073
9 (RC_{INDIV})	n.s.	45.9	61.4	31.5	74.6	57.1	1.4	0.067
10 (SRB_{MCS})	0.001	27.9	94.3	65.4	77.2	75.8	6.4	0.307
11 (SRB_{MULT})	0.001	26.2	93.7	61.5	76.7	74.9	5.3	0.276
12 (SRB_{CH})	0.001	27.9	94.3	65.4	77.2	75.8	6.4	0.307

Phase F: RAVLT Delayed Recall (n = 218; 157 no loss and 61 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	32.8	78.3	37.0	75.0	65.6	1.8	0.116
3 (RCI)	n.s.	32.8	78.3	37.0	75.0	65.6	1.8	0.116
4 (RCI_{SPEER})	0.05	24.6	87.3	42.9	74.9	69.7	2.2	0.145
5a ($RCI_{JT}a$)	n.s.	11.5	90.4	31.8	72.4	68.3	1.2	0.029
5b ($RCI_{JT}b$)	n.s.	32.8	78.3	37.0	75.0	65.6	1.8	0.116
5c ($RCI_{JT}c$)	n.s.	11.5	90.4	31.8	72.4	68.3	1.2	0.029
6a ($RCI_{CHEL}a$)	n.s.	8.2	94.9	38.5	72.7	70.6	1.7	0.059
6b ($RCI_{CHEL}b$)	n.s.	23.0	85.4	37.8	74.0	67.9	1.7	0.099
6c ($RCI_{CHEL}c$)	n.s.	8.2	94.9	38.5	72.7	70.6	1.7	0.059
7 (RCI_{HSU})	0.05	18.0	91.7	45.8	74.2	71.1	2.4	0.140
8 (RC_{ID})	n.s.	32.8	78.3	37.0	75.0	65.6	1.8	0.116
9 (RC_{INDIV})	n.s.	32.8	78.3	37.0	75.0	65.6	1.8	0.116
10 (SRB_{MCS})	n.s.	13.1	93.0	42.1	73.4	70.6	2.0	0.097
11 (SRB_{MULT})	n.s.	8.2	93.0	31.3	72.3	69.3	1.2	0.020
12 (SRB_{CH})	n.s.	13.1	93.0	42.1	73.4	70.6	2.0	0.097

Phase F: BVRT Multiple Choice (n = 234; 168 no loss and 66 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	25.8	82.1	36.2	73.8	66.2	1.6	0.089
3 (RCI)	n.s.	25.8	82.1	36.2	73.8	66.2	1.6	0.089
4 (RCI_{SPEER})	n.s.	19.7	88.1	39.4	73.6	68.8	1.8	0.101
5a ($RCI_{JT}a$)	0.05	18.2	91.1	44.4	73.9	70.5	2.3	0.130
5b ($RCI_{JT}b$)	n.s.	25.8	82.1	36.2	73.8	66.2	1.6	0.089
5c ($RCI_{JT}c$)	0.05	18.2	91.1	44.4	73.9	70.5	2.3	0.130
6a ($RCI_{CHEL}a$)	0.05	12.1	95.2	50.0	73.4	71.8	2.8	0.131
6b ($RCI_{CHEL}b$)	0.05	18.2	91.1	44.4	73.9	70.5	2.3	0.130
6c ($RCI_{CHEL}c$)	0.05	18.2	91.1	44.4	73.9	70.5	2.3	0.130
7 (RCI_{HSU})	n.s.	19.7	88.1	39.4	73.6	68.8	1.8	0.101
8 (RC_{ID})	-	0.0	100.0	0.0	71.8	71.8	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	24.2	85.7	40.0	74.2	68.4	1.9	0.119
11 (SRB_{MULT})	n.s.	22.7	87.5	41.7	74.2	69.2	2.1	0.128
12 (SRB_{CH})	n.s.	24.2	85.7	40.0	74.2	68.4	1.9	0.119

Phase F: WAIS-R Similarities subtest (n = 249; 176 no loss and 73 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	13.7	92.0	41.7	72.0	69.1	1.8	0.089
3 (RCI)	n.s.	16.4	85.2	31.6	71.1	65.1	1.1	0.021
4 (RCI_{SPEER})	n.s.	16.4	87.5	35.3	71.6	66.7	1.4	0.052
5a ($RCI_{JT}a$)	n.s.	13.7	92.0	41.7	72.0	69.1	1.8	0.089
5b ($RCI_{JT}b$)	n.s.	16.4	85.2	31.6	71.1	65.1	1.1	0.021
5c ($RCI_{JT}c$)	n.s.	13.7	92.0	41.7	72.0	69.1	1.8	0.089
6a ($RCI_{CHEL}a$)	n.s.	13.7	92.0	41.7	72.0	69.1	1.8	0.089
6b ($RCI_{CHEL}b$)	n.s.	16.4	85.2	31.6	71.1	65.1	1.1	0.021
6c ($RCI_{CHEL}c$)	n.s.	13.7	92.0	41.7	72.0	69.1	1.8	0.089
7 (RCI_{HSU})	n.s.	15.1	89.8	37.9	71.8	67.9	1.6	0.069
8 (RC_{ID})	0.01	9.6	98.3	70.0	72.4	72.3	6.1	0.183
9 (RC_{INDIV})	n.s.	13.7	92.0	41.7	72.0	69.1	1.8	0.089
10 (SRB_{MCS})	n.s.	15.1	88.1	34.4	71.4	66.7	1.3	0.043
11 (SRB_{MULT})	n.s.	13.7	91.5	40.0	71.9	68.7	1.7	0.078
12 (SRB_{CH})	n.s.	15.1	88.1	34.4	71.4	66.7	1.3	0.043

Phase F: WAIS-R Comprehension subtest (n = 249; 177 no loss and 72 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	16.7	89.3	38.7	72.5	68.3	1.7	0.081
3 (RCI)	n.s.	16.7	89.3	38.7	72.5	68.3	1.7	0.081
4 (RCI_{SPEER})	n.s.	22.2	87.0	41.0	73.3	68.3	1.9	0.115
5a ($RCI_{JT}a$)	0.05	13.9	93.8	47.6	72.8	70.7	2.4	0.125
5b ($RCI_{JT}b$)	n.s.	16.7	89.3	38.7	72.5	68.3	1.7	0.081
5c ($RCI_{JT}c$)	0.05	13.9	93.8	47.6	72.8	70.7	2.4	0.125
6a ($RCI_{CHEL}a$)	0.05	13.9	93.8	47.6	72.8	70.7	2.4	0.125
6b ($RCI_{CHEL}b$)	n.s.	16.7	89.3	38.7	72.5	68.3	1.7	0.081
6c ($RCI_{CHEL}c$)	0.05	13.9	93.8	47.6	72.8	70.7	2.4	0.125
7 (RCI_{HSU})	0.05	18.1	91.5	46.4	73.3	70.3	2.4	0.137
8 (RC_{ID})	-	0.0	100.0	0.0	71.1	71.1	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.05	16.7	92.7	48.0	73.2	70.7	2.5	0.141
11 (SRB_{MULT})	0.05	15.3	93.2	47.8	73.0	70.7	2.5	0.133
12 (SRB_{CH})	0.05	16.7	93.8	52.2	73.5	71.5	3.0	0.164

Phase F: Token Test (n = 238; 169 no loss and 69 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	23.2	82.2	34.8	72.4	65.1	1.4	0.062
3 (RCI)	n.s.	13.0	88.8	32.1	71.4	66.8	1.2	0.025
4 (RCI_{SPEER})	n.s.	14.5	88.2	33.3	71.6	66.8	1.3	0.036
5a ($RCI_{JT}a$)	n.s.	7.2	92.3	27.8	70.9	67.6	0.9	-0.008
5b ($RCI_{JT}b$)	n.s.	23.2	82.2	34.8	72.4	65.1	1.4	0.062
5c ($RCI_{JT}c$)	n.s.	13.0	88.8	32.1	71.4	66.8	1.2	0.025
6a ($RCI_{CHEL}a$)	n.s.	5.8	94.7	30.8	71.1	68.9	1.1	0.009
6b ($RCI_{CHEL}b$)	n.s.	23.2	82.2	34.8	72.4	65.1	1.4	0.062
6c ($RCI_{CHEL}c$)	n.s.	7.2	92.3	27.8	70.9	67.6	0.9	-0.008
7 (RCI_{HSU})	n.s.	11.6	90.5	33.3	71.5	67.6	1.3	0.032
8 (RC_{ID})	n.s.	23.2	82.2	34.8	72.4	65.1	1.4	0.062
9 (RC_{INDIV})	n.s.	13.0	88.8	32.1	71.4	66.8	1.2	0.025
10 (SRB_{MCS})	n.s.	17.4	88.2	37.5	72.3	67.6	1.6	0.074
11 (SRB_{MULT})	n.s.	17.4	86.4	34.3	71.9	66.4	1.3	0.048
12 (SRB_{CH})	n.s.	17.4	88.2	37.5	72.3	67.6	1.6	0.074

Phase F: Phonemic (FAS) fluency (n = 245; 173 no loss and 72 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	16.7	94.8	57.1	73.2	71.8	3.6	0.187
3 (RCI)	0.01	30.6	85.5	46.8	74.7	69.4	2.6	0.186
4 (RCI_{SPEER})	0.01	26.4	90.2	52.8	74.6	71.4	3.3	0.213
5a ($RCI_{JT}a$)	0.05	12.5	95.4	52.9	72.4	71.0	2.9	0.141
5b ($RCI_{JT}b$)	0.01	48.6	71.1	41.2	76.9	64.5	2.3	0.189
5c ($RCI_{JT}c$)	0.05	12.5	95.4	52.9	72.4	71.0	2.9	0.141
6a ($RCI_{CHEL}a$)	n.s.	6.9	97.1	50.0	71.5	70.6	2.5	0.093
6b ($RCI_{CHEL}b$)	0.01	44.4	75.7	43.2	76.6	66.5	2.5	0.200
6c ($RCI_{CHEL}c$)	0.05	12.5	95.4	52.9	72.4	71.0	2.9	0.141
7 (RCI_{HSU})	0.01	13.9	96.0	58.8	72.8	71.8	3.8	0.176
8 (RC_{ID})	0.05	12.5	95.4	52.9	72.4	71.0	2.9	0.141
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.05	13.9	94.8	52.6	72.6	71.0	2.9	0.148
11 (SRB_{MULT})	0.05	12.5	95.4	52.9	72.4	71.0	2.9	0.141
12 (SRB_{CH})	0.05	13.9	94.8	52.6	72.6	71.0	2.9	0.148

Phase F: Category (Animals) fluency (n = 245; 174 no loss and 71 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.05	29.6	83.3	42.0	74.4	67.8	2.1	0.145
3 (RCI)	0.05	29.6	83.3	42.0	74.4	67.8	2.1	0.145
4 (RCI_{SPEER})	0.05	25.4	87.9	46.2	74.3	69.8	2.5	0.165
5a ($RCI_{JT}a$)	0.05	16.9	92.0	46.2	73.1	70.2	2.3	0.130
5b ($RCI_{JT}b$)	0.01	39.4	78.7	43.1	76.1	67.3	2.4	0.187
5c ($RCI_{JT}c$)	0.05	16.9	92.0	46.2	73.1	70.2	2.3	0.130
6a ($RCI_{CHEL}a$)	0.05	14.1	94.3	50.0	72.9	71.0	2.7	0.138
6b ($RCI_{CHEL}b$)	0.05	29.6	83.3	42.0	74.4	67.8	2.1	0.145
6c ($RCI_{CHEL}c$)	0.05	14.1	94.3	50.0	72.9	71.0	2.7	0.138
7 (RCI_{HSU})	0.001	18.3	96.0	65.0	74.2	73.5	5.3	0.237
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	18.3	96.0	65.0	74.2	73.5	5.3	0.237
11 (SRB_{MULT})	0.01	14.1	96.6	62.5	73.4	72.7	4.6	0.195
12 (SRB_{CH})	0.001	18.3	96.0	65.0	74.2	73.5	5.3	0.237

Phase F: WAIS-R Block Design subtest (n = 240; 172 no loss and 68 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	29.4	72.7	29.9	72.3	60.4	1.1	0.021
3 (RCI)	n.s.	19.1	80.2	27.7	71.5	62.9	1.0	-0.007
4 (RCI_{SPEER})	0.05	26.5	86.0	42.9	74.7	69.2	2.2	0.148
5a ($RCI_{JT}a$)	n.s.	14.7	88.4	33.3	72.4	67.5	1.3	0.042
5b ($RCI_{JT}b$)	n.s.	42.6	61.6	30.5	73.1	56.3	1.2	0.039
5c ($RCI_{JT}c$)	n.s.	16.2	86.0	31.4	72.2	66.3	1.2	0.028
6a ($RCI_{CHEL}a$)	n.s.	2.9	92.4	13.3	70.7	67.1	0.4	-0.086
6b ($RCI_{CHEL}b$)	n.s.	19.1	80.2	27.7	71.5	62.9	1.0	-0.007
6c ($RCI_{CHEL}c$)	n.s.	8.8	90.7	27.3	71.6	67.5	1.0	-0.007
7 (RCI_{HSU})	n.s.	10.3	95.3	46.7	72.9	71.3	2.4	0.105
8 (RC_{ID})	-	0.0	100.0	0.0	71.7	71.7	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	14.7	90.7	38.5	72.9	69.2	1.7	0.078
11 (SRB_{MULT})	0.05	14.7	94.2	50.0	73.6	71.7	2.8	0.145
12 (SRB_{CH})	n.s.	14.7	91.3	40.0	73.0	69.6	1.8	0.088

Phase F: WAIS-R Digit Symbol subtest (n = 220; 160 no loss and 60 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	33.3	92.5	62.5	78.7	76.4	6.2	0.326
3 (RCI)	0.001	55.0	70.6	41.3	80.7	66.4	2.9	0.237
4 (RCI_{SPEER})	0.001	53.3	72.5	42.1	80.6	67.3	3.0	0.242
5a ($RCI_{JT}a$)	0.001	35.0	86.3	48.8	78.0	72.3	3.4	0.239
5b ($RCI_{JT}b$)	0.05	66.7	51.3	33.9	80.4	55.5	2.1	0.160
5c ($RCI_{JT}c$)	0.001	35.0	86.3	48.8	78.0	72.3	3.4	0.239
6a ($RCI_{CHEL}a$)	0.001	20.0	95.6	63.2	76.1	75.0	5.5	0.248
6b ($RCI_{CHEL}b$)	0.01	40.0	80.0	42.9	78.0	69.1	2.7	0.204
6c ($RCI_{CHEL}c$)	0.05	11.7	96.3	53.8	74.4	73.2	3.4	0.150
7 (RCI_{HSU})	0.001	25.0	95.0	65.2	77.2	75.9	6.3	0.291
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.01	16.7	95.6	58.8	75.4	74.1	4.4	0.205
11 (SRB_{MULT})	0.001	18.3	96.9	68.8	76.0	75.5	7.0	0.261
12 (SRB_{CH})	0.01	16.7	95.6	58.8	75.4	74.1	4.4	0.205

Phase G: Wechsler Memory Scale - Information (n = 224; 173 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	62.7	61.3	32.3	82.8	61.6	2.7	0.203
3 (RCI)	0.001	41.2	86.1	46.7	83.2	75.9	4.3	0.286
4 (RCI_{SPEER})	0.001	51.0	80.9	44.1	84.8	74.1	4.4	0.304
5a ($RCI_{JT}a$)	0.001	41.2	86.1	46.7	83.2	75.9	4.3	0.286
5b ($RCI_{JT}b$)	0.001	19.6	96.5	62.5	80.3	79.0	6.8	0.263
5c ($RCI_{JT}c$)	0.001	41.2	86.1	46.7	83.2	75.9	4.3	0.286
6a ($RCI_{CHEL}a$)	0.001	41.2	86.1	46.7	83.2	75.9	4.3	0.286
6b ($RCI_{CHEL}b$)	0.001	19.6	96.5	62.5	80.3	79.0	6.8	0.263
6c ($RCI_{CHEL}c$)	0.001	19.6	96.5	62.5	80.3	79.0	6.8	0.263
7 (RCI_{HSU})	0.001	39.2	88.4	50.0	83.2	77.2	4.9	0.303
8 (RC_{ID})	0.001	11.8	99.4	85.7	79.3	79.5	22.9	0.270
9 (RC_{INDIV})	0.001	41.2	86.1	46.7	83.2	75.9	4.3	0.286
10 (SRB_{MCS})	0.001	29.4	93.1	55.6	89.7	78.6	5.6	0.289
11 (SRB_{MULT})	0.001	25.5	92.5	50.0	80.8	77.2	4.2	0.235
12 (SRB_{CH})	0.001	29.4	93.1	55.6	89.7	78.6	5.6	0.289

Phase G: Buschke Cued Recall – Free Recall Trial 1 (n = 218; 167 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	49.0	77.8	40.3	83.3	71.1	3.4	0.252
3 (RCI)	0.001	49.0	77.8	40.3	83.3	71.1	3.4	0.252
4 (RCI_{SPEER})	0.01	33.3	86.2	42.5	80.9	73.9	3.1	0.214
5a ($RCI_{JT}a$)	0.001	37.3	90.4	54.3	82.5	78.0	5.6	0.319
5b ($RCI_{JT}b$)	0.001	37.3	90.4	54.3	82.5	78.0	5.6	0.319
5c ($RCI_{JT}c$)	0.001	37.3	90.4	54.3	82.5	78.0	5.6	0.319
6a ($RCI_{CHEL}a$)	0.001	21.6	96.4	64.7	80.1	78.9	7.4	0.284
6b ($RCI_{CHEL}b$)	0.001	21.6	96.4	64.7	80.1	78.9	7.4	0.284
6c ($RCI_{CHEL}c$)	0.001	21.6	96.4	64.7	80.1	78.9	7.4	0.284
7 (RCI_{HSU})	0.001	29.4	92.2	53.6	81.1	77.5	4.9	0.274
8 (RC_{ID})	-	0.0	100.0	0.0	76.6	76.6	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	27.5	94.0	58.3	80.9	78.4	5.9	0.290
11 (SRB_{MULT})	0.001	23.5	95.8	63.2	80.4	78.9	7.0	0.290
12 (SRB_{CH})	0.001	27.5	94.0	58.3	80.9	78.4	5.9	0.290

Phase G: Buschke Retrieval (n = 218; 167 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	62.7	80.2	49.2	87.6	76.1	6.8	0.398
3 (RCI)	0.001	62.7	80.2	49.2	87.6	76.1	6.8	0.398
4 (RCI_{SPEER})	0.001	56.9	83.8	51.8	86.4	77.5	6.8	0.394
5a ($RCI_{JT}a$)	0.001	47.1	89.2	57.1	84.7	79.4	7.4	0.389
5b ($RCI_{JT}b$)	0.001	62.7	80.2	49.2	87.6	76.1	6.8	0.398
5c ($RCI_{JT}c$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
6a ($RCI_{CHEL}a$)	0.001	35.3	94.0	64.3	82.6	80.3	8.6	0.371
6b ($RCI_{CHEL}b$)	0.001	47.1	89.2	57.1	84.7	79.4	7.4	0.389
6c ($RCI_{CHEL}c$)	0.001	35.3	94.0	64.3	82.6	80.3	8.6	0.371
7 (RCI_{HSU})	0.001	43.1	91.6	61.1	84.1	80.3	8.3	0.396
8 (RC_{ID})	0.001	47.1	89.2	57.1	84.7	79.4	7.4	0.389
9 (RC_{INDIV})	0.001	52.9	84.4	50.9	85.5	77.1	6.1	0.369
10 (SRB_{MCS})	0.001	35.3	94.6	66.7	82.7	80.7	9.6	0.384
11 (SRB_{MULT})	0.001	37.3	93.4	63.3	83.0	80.3	8.4	0.377
12 (SRB_{CH})	0.001	35.3	94.6	66.7	82.7	80.7	9.6	0.384

Phase G: Buschke Acquisition (n = 218; 167 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	29.4	90.4	48.4	80.7	76.1	3.9	0.240
3 (RCI)	0.001	19.6	97.6	71.4	79.9	79.4	9.9	0.297
4 (RCI_{SPEER})	0.001	21.6	97.6	73.3	80.3	79.8	11.2	0.321
5a ($RCI_{JT}a$)	0.001	19.6	97.6	71.4	79.9	79.4	9.9	0.297
5b ($RCI_{JT}b$)	0.001	13.7	98.8	77.8	78.9	78.9	13.1	0.267
5c ($RCI_{JT}c$)	0.001	19.6	97.6	71.4	79.9	79.4	9.9	0.297
6a ($RCI_{CHEL}a$)	0.001	19.6	97.6	71.4	79.9	79.4	9.9	0.297
6b ($RCI_{CHEL}b$)	0.001	13.7	98.8	77.8	78.9	78.9	13.1	0.267
6c ($RCI_{CHEL}c$)	0.001	13.7	98.8	77.8	78.9	78.9	13.1	0.267
7 (RCI_{HSU})	0.001	21.6	97.6	73.3	80.3	79.8	11.2	0.321
8 (RC_{ID})	0.001	19.6	97.6	71.4	79.9	79.4	9.9	0.297
9 (RC_{INDIV})	0.001	29.4	90.4	48.4	80.7	76.1	3.9	0.240
10 (SRB_{MCS})	0.001	35.3	88.6	48.6	81.8	76.1	4.2	0.270
11 (SRB_{MULT})	0.001	35.3	88.6	48.6	81.8	76.1	4.2	0.270
12 (SRB_{CH})	0.001	35.3	88.6	48.6	81.8	76.1	4.2	0.270

Phase G: Buschke Delayed Free Recall (n = 218; 167 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.001	52.9	80.8	45.8	84.9	74.3	4.7	0.322
3 (RCI)	0.001	52.9	80.8	45.8	84.9	74.3	4.7	0.322
4 (RCI_{SPEER})	0.001	51.0	88.0	56.5	85.5	79.4	7.6	0.405
5a ($RCI_{JT}a$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
5b ($RCI_{JT}b$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
5c ($RCI_{JT}c$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
6a ($RCI_{CHEL}a$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
6b ($RCI_{CHEL}b$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
6c ($RCI_{CHEL}c$)	0.001	39.2	92.8	62.5	83.3	80.3	8.3	0.383
7 (RCI_{HSU})	0.001	47.1	91.6	63.2	85.0	81.2	9.7	0.432
8 (RC_{ID})	-	0.0	100.0	0.0	76.6	76.6	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	51.0	88.0	56.5	85.5	79.4	7.6	0.405
11 (SRB_{MULT})	0.001	51.0	89.2	59.1	85.6	80.3	8.6	0.424
12 (SRB_{CH})	0.001	51.0	88.0	56.5	85.5	79.4	7.6	0.405

Phase G: RAVLT Trial 1 (n = 208; 159 no loss and 49 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	34.7	78.6	33.3	79.6	68.3	2.0	0.131
3 (RCI)	n.s.	18.4	89.9	36.0	78.1	73.1	2.0	0.108
4 (RCI_{SPEER})	0.05	18.4	93.7	47.4	78.8	76.0	3.4	0.178
5a ($RCI_{JT}a$)	n.s.	10.2	95.0	38.5	77.4	75.0	2.1	0.091
5b ($RCI_{JT}b$)	n.s.	18.4	89.9	36.0	78.1	73.1	2.0	0.108
5c ($RCI_{JT}c$)	n.s.	10.2	95.0	38.5	77.4	75.0	2.1	0.091
6a ($RCI_{CHEL}a$)	n.s.	10.2	95.0	38.5	77.4	75.0	2.1	0.091
6b ($RCI_{CHEL}b$)	n.s.	18.4	89.9	36.0	78.1	73.1	2.0	0.108
6c ($RCI_{CHEL}c$)	n.s.	10.2	95.0	38.5	77.4	75.0	2.1	0.091
7 (RCI_{HSU})	0.05	12.2	96.2	50.0	78.1	76.4	3.6	0.154
8 (RC_{ID})	-	0.0	100.0	0.0	76.4	76.4	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.01	16.3	95.0	50.0	78.6	76.4	3.7	0.180
11 (SRB_{MULT})	0.05	14.3	95.0	46.7	78.2	76.0	3.1	0.152
12 (SRB_{CH})	0.01	16.3	95.0	50.0	78.6	76.4	3.7	0.180

Phase G: RAVLT Total score (n = 200; 151 no loss and 49 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	38.8	81.5	40.4	80.4	71.0	2.8	0.205
3 (RCI)	0.01	38.8	81.5	40.4	80.4	71.0	2.8	0.205
4 (RCI_{SPEER})	0.001	38.8	89.4	54.3	81.8	77.0	5.3	0.319
5a ($RCI_{JT}a$)	0.001	28.6	91.4	51.9	79.8	76.0	4.2	0.251
5b ($RCI_{JT}b$)	n.s.	53.1	60.3	30.2	79.8	58.5	1.7	0.116
5c ($RCI_{JT}c$)	0.001	28.6	91.4	51.9	79.8	76.0	4.2	0.251
6a ($RCI_{CHEL}a$)	0.001	24.5	96.0	66.7	79.7	78.5	7.8	0.308
6b ($RCI_{CHEL}b$)	0.05	40.8	75.5	35.1	79.7	67.0	2.1	0.155
6c ($RCI_{CHEL}c$)	0.001	24.5	96.0	66.7	79.7	78.5	7.8	0.308
7 (RCI_{HSU})	0.001	24.5	96.7	70.6	79.8	79.0	9.5	0.327
8 (RC_{ID})	n.s.	53.1	58.3	29.2	79.3	57.0	1.6	0.098
9 (RC_{INDIV})	0.05	51.0	65.6	32.5	80.5	62.0	2.0	0.147
10 (SRB_{MCS})	0.001	24.5	96.7	70.6	79.8	79.0	9.5	0.327
11 (SRB_{MULT})	0.001	28.6	97.4	77.8	80.8	80.5	14.7	0.390
12 (SRB_{CH})	0.001	24.5	96.7	70.6	79.8	79.0	9.5	0.327

Phase G: RAVLT Delayed Recall (n = 199; 151 no loss and 48 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	29.2	78.8	30.4	77.8	66.8	1.5	0.081
3 (RCI)	n.s.	29.2	78.8	30.4	77.8	66.8	1.5	0.081
4 (RCI_{SPEER})	0.05	20.8	90.1	40.0	78.2	73.4	2.4	0.141
5a ($RCI_{JT}a$)	n.s.	14.6	92.1	36.8	77.2	73.4	2.0	0.097
5b ($RCI_{JT}b$)	n.s.	29.2	78.8	30.4	77.8	66.8	1.5	0.081
5c ($RCI_{JT}c$)	n.s.	14.6	92.1	36.8	77.2	73.4	2.0	0.097
6a ($RCI_{CHEL}a$)	n.s.	10.4	96.0	45.5	77.1	75.4	2.8	0.121
6b ($RCI_{CHEL}b$)	n.s.	22.9	85.4	33.3	77.7	70.4	1.7	0.096
6c ($RCI_{CHEL}c$)	n.s.	10.4	96.0	45.5	77.1	75.4	2.8	0.121
7 (RCI_{HSU})	0.05	18.8	93.4	47.4	78.3	75.4	3.3	0.177
8 (RC_{ID})	n.s.	29.2	78.8	30.4	77.8	66.8	1.5	0.081
9 (RC_{INDIV})	n.s.	29.2	78.8	30.4	77.8	66.8	1.5	0.081
10 (SRB_{MCS})	n.s.	12.5	94.7	42.9	77.3	74.9	2.6	0.120
11 (SRB_{MULT})	n.s.	10.4	96.0	45.5	77.1	75.4	2.8	0.121
12 (SRB_{CH})	n.s.	12.5	94.7	42.9	77.3	74.9	2.6	0.120

Phase G: BVRT Multiple Choice (n = 208; 161 no loss and 47 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.05	29.8	86.3	38.9	80.8	73.6	2.7	0.178
3 (RCI)	0.05	29.8	86.3	38.9	80.8	73.6	2.7	0.178
4 (RCI_{SPEER})	0.01	25.5	92.5	50.0	81.0	77.4	4.3	0.237
5a ($RCI_{JT}a$)	0.05	17.0	94.4	47.1	79.6	76.9	3.5	0.175
5b ($RCI_{JT}b$)	0.05	29.8	86.3	38.9	80.8	73.6	2.7	0.178
5c ($RCI_{JT}c$)	0.05	17.0	94.4	47.1	79.6	76.9	3.5	0.175
6a ($RCI_{CHEL}a$)	n.s.	8.5	97.5	50.0	78.5	77.4	3.7	0.131
6b ($RCI_{CHEL}b$)	0.05	17.0	94.4	47.1	79.6	76.9	3.5	0.175
6c ($RCI_{CHEL}c$)	0.05	17.0	94.4	47.1	79.6	76.9	3.5	0.175
7 (RCI_{HSU})	0.01	25.5	92.5	50.0	81.0	77.4	4.3	0.237
8 (RC_{ID})	-	0.0	100.0	0.0	77.4	77.4	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.001	31.9	91.9	53.6	82.2	78.4	5.3	0.292
11 (SRB_{MULT})	0.001	27.7	92.5	52.0	81.4	77.9	4.7	0.260
12 (SRB_{CH})	0.001	31.9	91.9	53.6	82.2	78.4	5.3	0.292

Phase G: WAIS-R Similarities subtest (n = 224; 173 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	0.01	19.6	93.6	47.6	79.8	76.8	3.6	0.191
3 (RCI)	n.s.	19.6	87.3	31.3	78.6	71.9	1.7	0.083
4 (RCI_{SPEER})	0.05	21.6	89.6	37.9	79.5	74.1	2.4	0.139
5a ($RCI_{JT}a$)	0.01	19.6	93.6	47.6	79.8	76.8	3.6	0.191
5b ($RCI_{JT}b$)	n.s.	19.6	87.3	31.3	78.6	71.9	1.7	0.083
5c ($RCI_{JT}c$)	0.01	19.6	93.6	47.6	79.8	76.8	3.6	0.191
6a ($RCI_{CHEL}a$)	0.01	19.6	93.6	47.6	79.8	76.8	3.6	0.191
6b ($RCI_{CHEL}b$)	n.s.	19.6	87.3	31.3	78.6	71.9	1.7	0.083
6c ($RCI_{CHEL}c$)	0.01	19.6	93.6	47.6	79.8	76.8	3.6	0.191
7 (RCI_{HSU})	0.05	19.6	91.9	41.7	79.5	75.4	2.8	0.156
8 (RC_{ID})	0.001	11.8	98.8	75.0	79.2	79.0	11.4	0.240
9 (RC_{INDIV})	0.01	19.6	93.6	47.6	79.8	76.8	3.6	0.191
10 (SRB_{MCS})	0.05	21.6	90.8	40.7	79.7	75.0	2.7	0.159
11 (SRB_{MULT})	0.05	17.6	92.5	40.9	79.2	75.4	2.6	0.143
12 (SRB_{CH})	0.05	21.6	90.8	40.7	79.7	75.0	2.7	0.159

Phase G: WAIS-R Comprehension subtest (n = 223; 172 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	17.6	90.7	36.0	78.8	74.0	2.1	0.111
3 (RCI)	n.s.	17.6	90.7	36.0	78.8	74.0	2.1	0.111
4 (RCI_{SPEER})	n.s.	21.6	88.4	35.5	79.2	73.1	2.1	0.121
5a ($RCI_{JT}a$)	n.s.	9.8	95.3	38.5	78.1	75.8	2.2	0.092
5b ($RCI_{JT}b$)	n.s.	17.6	90.7	36.0	78.8	74.0	2.1	0.111
5c ($RCI_{JT}c$)	n.s.	9.8	95.3	38.5	78.1	75.8	2.2	0.092
6a ($RCI_{CHEL}a$)	n.s.	9.8	95.3	38.5	78.1	75.8	2.2	0.092
6b ($RCI_{CHEL}b$)	n.s.	17.6	90.7	36.0	78.8	74.0	2.1	0.111
6c ($RCI_{CHEL}c$)	n.s.	9.8	95.3	38.5	78.1	75.8	2.2	0.092
7 (RCI_{HSU})	n.s.	15.7	93.0	40.0	78.8	75.3	2.5	0.123
8 (RC_{ID})	-	0.0	100.0	0.0	77.1	77.1	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	13.7	93.6	38.9	78.5	75.3	2.3	0.113
11 (SRB_{MULT})	n.s.	11.8	94.2	37.5	78.3	75.3	2.1	0.097
12 (SRB_{CH})	0.05	13.7	94.8	43.8	78.7	76.2	2.9	0.138

Phase G: Token Test (n = 212; 164 no loss and 48 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	18.8	86.0	28.1	78.3	70.8	1.4	0.055
3 (RCI)	n.s.	12.5	91.5	30.0	78.1	73.6	1.5	0.057
4 (RCI_{SPEER})	n.s.	14.6	90.2	30.4	78.3	73.1	1.6	0.065
5a ($RCI_{JT}a$)	n.s.	6.3	94.5	25.0	77.5	74.5	1.1	0.014
5b ($RCI_{JT}b$)	n.s.	18.8	86.0	28.1	78.3	70.8	1.4	0.055
5c ($RCI_{JT}c$)	n.s.	12.5	91.5	30.0	78.1	73.6	1.5	0.057
6a ($RCI_{CHEL}a$)	n.s.	4.2	97.0	28.6	77.6	75.9	1.4	0.026
6b ($RCI_{CHEL}b$)	n.s.	18.8	86.0	28.1	78.3	70.8	1.4	0.055
6c ($RCI_{CHEL}c$)	n.s.	6.3	94.5	25.0	77.5	74.5	1.1	0.014
7 (RCI_{HSU})	n.s.	10.4	93.3	31.3	78.1	74.5	1.6	0.059
8 (RC_{ID})	n.s.	18.8	86.0	28.1	78.3	70.8	1.4	0.055
9 (RC_{INDIV})	n.s.	12.5	91.5	30.0	78.1	73.6	1.5	0.057
10 (SRB_{MCS})	n.s.	16.7	90.2	33.3	78.7	73.6	1.9	0.091
11 (SRB_{MULT})	n.s.	16.7	88.4	29.6	78.4	72.2	1.5	0.064
12 (SRB_{CH})	n.s.	16.7	90.2	33.3	78.7	73.6	1.9	0.091

Phase G: Phonemic (FAS) fluency (n = 218; 167 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	5.9	94.6	25.0	76.7	73.9	1.1	0.009
3 (RCI)	0.05	25.5	86.8	37.1	79.2	72.5	2.3	0.142
4 (RCI_{SPEER})	n.s.	17.6	91.0	37.5	78.4	73.9	2.1	0.117
5a ($RCI_{JT}a$)	n.s.	3.9	95.8	22.2	76.6	74.3	0.9	-0.006
5b ($RCI_{JT}b$)	n.s.	41.2	72.5	31.3	80.1	65.1	1.8	0.125
5c ($RCI_{JT}c$)	n.s.	3.9	95.8	22.2	76.6	74.3	0.9	-0.006
6a ($RCI_{CHEL}a$)	n.s.	0.0	97.6	0.0	76.2	74.8	1.0	-0.076
6b ($RCI_{CHEL}b$)	0.05	39.2	76.6	33.9	80.5	67.9	2.1	0.151
6c ($RCI_{CHEL}c$)	n.s.	3.9	95.8	22.2	76.6	74.3	0.9	-0.006
7 (RCI_{HSU})	n.s.	5.9	96.4	33.3	77.0	75.2	1.7	0.049
8 (RC_{ID})	n.s.	3.9	95.8	22.2	76.6	74.3	0.9	-0.006
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	7.8	95.8	36.4	77.3	75.2	1.9	0.071
11 (SRB_{MULT})	n.s.	5.9	95.8	30.0	76.9	74.8	1.4	0.034
12 (SRB_{CH})	n.s.	7.8	95.8	36.4	77.3	75.2	1.9	0.071

Phase G: Category (Animals) fluency (n = 220; 169 no loss and 51 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	23.5	84.0	30.8	78.5	70.0	1.6	0.083
3 (RCI)	n.s.	23.5	84.0	30.8	78.5	70.0	1.6	0.083
4 (RCI_{SPEER})	0.01	25.5	92.3	50.0	80.4	76.8	4.1	0.233
5a ($RCI_{JT}a$)	n.s.	11.8	92.3	31.6	77.6	73.6	1.6	0.061
5b ($RCI_{JT}b$)	n.s.	29.4	79.3	30.0	78.8	67.7	1.6	0.088
5c ($RCI_{JT}c$)	n.s.	11.8	92.3	31.6	77.6	73.6	1.6	0.061
6a ($RCI_{CHEL}a$)	n.s.	9.8	94.1	33.3	77.6	74.5	1.7	0.065
6b ($RCI_{CHEL}b$)	n.s.	23.5	84.0	30.8	78.5	70.0	1.6	0.083
6c ($RCI_{CHEL}c$)	n.s.	9.8	94.1	33.3	77.6	74.5	1.7	0.065
7 (RCI_{HSU})	0.01	11.8	97.6	60.0	78.6	77.7	5.5	0.190
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	0.01	11.8	97.6	60.0	78.6	77.7	5.5	0.190
11 (SRB_{MULT})	0.05	7.8	98.2	57.1	77.9	77.3	4.7	0.146
12 (SRB_{CH})	0.01	11.8	97.6	60.0	78.6	77.7	5.5	0.190

Phase G: WAIS-R Block Design subtest (n = 214; 166 no loss and 48 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	27.1	72.9	22.4	77.6	72.6	1.0	0.000
3 (RCI)	n.s.	14.6	79.5	17.1	76.3	65.0	0.7	-0.063
4 (RCI_{SPEER})	n.s.	16.7	86.7	26.7	78.3	71.0	1.3	0.041
5a ($RCI_{JT}a$)	n.s.	6.3	87.3	12.5	76.3	69.2	0.5	-0.085
5b ($RCI_{JT}b$)	n.s.	45.8	64.5	27.2	80.5	60.3	1.5	0.089
5c ($RCI_{JT}c$)	n.s.	8.3	84.9	13.8	76.2	67.8	0.5	-0.082
6a ($RCI_{CHEL}a$)	n.s.	0.0	92.8	0.0	76.2	72.0	1.0	-0.131
6b ($RCI_{CHEL}b$)	n.s.	14.6	79.5	17.1	76.3	65.0	0.7	-0.063
6c ($RCI_{CHEL}c$)	n.s.	2.1	90.4	5.9	76.1	70.6	0.2	-0.117
7 (RCI_{HSU})	n.s.	6.3	95.2	27.3	77.8	75.2	1.3	0.027
8 (RC_{ID})	-	0.0	100.0	0.0	77.6	77.6	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	14.6	91.0	31.8	78.6	73.8	1.7	0.076
11 (SRB_{MULT})	n.s.	10.4	93.4	31.3	78.3	74.8	1.6	0.060
12 (SRB_{CH})	n.s.	14.6	91.6	33.3	78.8	74.3	1.9	0.086

Phase G: WAIS-R Digit Symbol subtest (n = 199; 156 no loss and 43 loss)

Equation	Chi^2	SENS	SPEC	PPV	NPV	CA	OR	r
2 (SD)	n.s.	14.0	90.4	28.6	79.2	73.9	1.5	0.058
3 (RCI)	n.s.	32.6	67.9	21.9	78.5	60.3	1.0	0.004
4 (RCI_{SPEER})	n.s.	34.9	72.4	25.9	80.1	64.3	1.4	0.066
5a ($RCI_{JT}a$)	n.s.	16.3	84.0	21.9	78.4	69.3	1.0	0.003
5b ($RCI_{JT}b$)	n.s.	51.2	48.7	21.6	78.4	49.2	1.0	-0.001
5c ($RCI_{JT}c$)	n.s.	16.3	84.0	21.9	78.4	69.3	1.0	0.003
6a ($RCI_{CHEL}a$)	n.s.	9.3	94.9	33.3	79.1	76.4	1.9	0.072
6b ($RCI_{CHEL}b$)	n.s.	18.6	78.2	19.0	77.7	65.3	0.8	-0.032
6c ($RCI_{CHEL}c$)	n.s.	4.7	96.8	28.6	78.6	76.9	1.5	0.032
7 (RCI_{HSU})	n.s.	11.6	94.2	35.7	79.5	76.4	2.1	0.094
8 (RC_{ID})	-	-	-	-	-	-	-	-
9 (RC_{INDIV})	-	-	-	-	-	-	-	-
10 (SRB_{MCS})	n.s.	9.3	96.2	40.0	79.4	77.4	2.6	0.103
11 (SRB_{MULT})	n.s.	7.0	95.5	30.0	78.8	76.4	1.6	0.047
12 (SRB_{CH})	n.s.	9.3	96.2	40.0	79.4	77.4	2.6	0.103