



The Assessment of Change: Serial Assessments in Dementia Evaluations

5

Gordon J. Chelune and Kevin Duff

The Assessment of Change: Serial Assessments in Dementia Evaluations

As a multidisciplinary area of scientific inquiry, *neuropsychology* is often defined as the study of brain–behavior relationships. However, as an area of psychological practice, *clinical neuropsychology* has been described as the application of neuropsychological principles of brain–behavior relationships to the assessment, diagnosis, and rehabilitation of *changes* in human behavior that arise across the lifespan from known or suspected illnesses or injuries affecting the brain [1]. To this definition, we can also add the assessment of cognitive changes associated with medical interventions (e.g., open-heart surgery, epilepsy surgery) and treatments (e.g., deep brain stimulation, pharmacologic treatments). Whether the focus is on changes in cognition induced by abnormal medical conditions or those in response to treatments and interventions, the focus of the clinical neuropsychologist in everyday practice is on *change*.

The assessment of meaningful neurocognitive change is particularly relevant for the evaluation

of older adults suspected of having underlying neurodegenerative disorders. Because the diagnosis of dementia as well as mild cognitive impairment (MCI) requires evidence of cognitive decline over time [2], it is critical to distinguish between age-related decrements in cognition (e.g., memory, processing speed, executive functions) believed to be part of “normal” aging [3–5] and those early clinical changes that are pathological and disease-related (e.g., neurodegenerative disorders, cerebrovascular disease, stroke, diabetes, etc.). Traditional single-point evaluations are limited in this context as they only capture a picture of the patient’s current abilities at a single point in time. Unless the patient’s performances deviate markedly from an *inferred* pre-morbid baseline, it is difficult for the practitioner to know whether these point estimates of a patient’s abilities are meaningfully different from expectation [6]. To overcome the limitations of single-point assessments, clinicians increasingly are turning to serial assessments to determine whether patients’ *observed* trajectories of change over time significantly deviate from those seen in normal aging [7, 8]. Unlike single-point assessments where the clinician must infer a pre-morbid baseline, the patient’s initial scores serve as their *observed* baseline. Armed with an appropriate conceptual framework and some simple tools, serial assessments provide the informed practitioner a powerful means for assessing diagnostically meaningful change.

G. J. Chelune · K. Duff (✉)

Department of Neurology, Center for Alzheimer’s Care, Imaging and Research, University of Utah School of Medicine, Salt Lake City, UT, USA
e-mail: gordon.chelune@hsc.utah.edu; kevin.duff@hsc.utah.edu

In this chapter, we will briefly discuss the clinical use of norm-referenced neuropsychological tests, contrasting two underlying approaches to interpreting these norms in traditional single-point assessments. With this as a backdrop, we will then turn our attention to the use of serial assessments to objectively monitor and assess cognitive changes over time, discussing the unique advantages and challenges of serial assessments. An overview and distillation of reliable change methods will be presented and applied to a case example, demonstrating how these methods can be used as effective tools to inform the clinical evaluation of the individual patient. In the end, we hope to leave the reader with an appreciation that *change* is a unique variable with its own inherent statistical properties and clinical meaning.

Norms and How We Use Them in Single-Point Assessments

In clinical practice, when we see a patient for the first time, we use norm-referenced tests so that we can compare the performances of the individual patient to an external reference group. The norms simply describe the distribution of scores on a given test obtained by a reference group, which can be a sample from the general population, a well-screened group of healthy community-living individuals (i.e., robust norms), or a patient group with a specific condition of interest. To infer meaning from our patient's scores, we can take two very distinct approaches to answer different clinical questions [6]. The first approach is *descriptive*, that is, *where* does my patient's score fall with respect to the reference population along a standardized metric (e.g., standard scores, z-scores, percentile ranks)? We often apply descriptive labels such as "above average" or "below average" for ranges of scores in relation to the mean of the sample, and using standardized measures of the distribution of scores, we can assign percentile ranks that tell us how common or uncommon the specific score is *within* the reference population.

While the descriptive approach is useful in identifying where our patient's scores fall within a reference population, it does not address whether our patient's scores are impaired or not. To do this, we must take a *diagnostic* approach where we ask the question "does my patient's score *deviate* from premorbid expectations (i.e., where I expect the score to have been in the absence of an intervening illness or injury), and if so, by how much?" The reference standard is now the individual's premorbid status, *not* the mean of the reference population. In the absence of having baseline information, the clinician must infer this and often relies on demographic information [9] and performance on crystallized ability measures such as oral reading derived from normative reference groups (e.g., the Test of Premorbid Functioning [10]). Deviations from this *individual comparison standard* can also be placed on a standardized metric (e.g., *T*-scores, *z*-scores), and percentile ranks assigned to the deviations *if* we know the characteristics of the *distribution of the deviation scores* between the premorbid estimate and observed performance on a given test. Note that the focus is on the distribution of the deviation scores, *not* the distribution of either the premorbid estimates or the observed scores on a given test.

While the diagnostic approach allows us to quantify whether an individual's current performance deviates from estimates of his or her demographically predicted premorbid ability level, we are still constrained to describing the deviation in terms of base rates—how common or uncommon the deviation is for our patient relative to premorbid expectations. To be diagnostically useful, the clinician must further establish validity evidence. As neuropsychologists move more concertedly toward evidence-based practice [11], it is no longer sufficient to simply rely on personal case records, unsystematic observations, or general knowledge as validity [12]. Increasingly, clinicians must become skilled in performing evidence-based reviews of the literature [13] that allow the integration of "...best research derived from the study of populations to inform clinical decisions about individuals within the context of the provider's expertise and individual patient values with the goal of maximizing clinical outcomes and quality of life..." (Chelune, 2017,

p 160). Our interpretation that discrepancies of a certain magnitude are statistically more frequent in populations that have a specific condition of interest, such as amnesic MCI, than would be expected at this level of discrepancy in a normal population should be founded on empirical evidence.

To illustrate the points above, let us consider the example of super clinician, Dr. Bob, who works in a memory disorders clinic and uses the test MegaMemory to evaluate memory complaints. Knowing that a patient's memory score on MegaMemory is one standard deviation below the estimated premorbid level informs Dr. Bob that the base rate of deviations of this magnitude occurs in only 16% of cases where there is an absence of an intervening illness or injury. However, after carefully reading the chapter on validity in the test manual for MegaMemory, Dr. Bob finds that the publisher conducted a case-controlled study using MegaMemory that compared equal numbers of patients with amnesic MCI and normal controls, a prevalence rate similar to what Dr. Bob sees in his clinic. The manual reports that individual deviations of one standard deviation or more from estimated premorbid levels occurred in 64% of cases with amnesic MCI compared to only 16% of controls. Performing a Bayesian analysis of the base rates between the two groups [13] yielded an odds ratio of 9.3 and a likelihood ratio of 4.0. Based on this empirical evidence, Dr. Bob now feels he can interpret a deviation score of one standard deviation or more on MegaMemory as not only relatively uncommon among healthy older adults but also as being "impaired" since deviations of this magnitude are four times more likely to occur in patients with amnesic MCI than in healthy controls, and among patients with amnesic MCI, deviations of this magnitude are nine times more likely to occur than deviations of lesser magnitude.

Using Serial Assessments to Identify Meaningful Change

Although neuropsychological tests are generally designed to assess the current state or capacity of an individual, repeated assessments are increas-

ingly common in neuropsychological practice and outcomes research [14, 15]. This has become especially true in geriatric settings where the determination of meaningful changes in cognition over time is essential for both the diagnosis of dementia and for planning therapeutic provisions and long-term care for patients and caregivers [6, 16]. Serial observations and longitudinal comparisons are classic tools in science, and their use in clinical practice requires clinicians to understand test–retest change scores as unique cognitive variables with their own statistical and clinical properties that are different from the test measures from which they were derived [17].

Like single-point diagnostic assessments discussed above, serial assessments share (a) a focus on *change* between two points in time (albeit one observed and the other inferred); (b) estimates of change based on *individual comparison standards* rather than population standards; (c) a focus on the psychometric properties of the discrepancy or change scores rather than on the test scores themselves (i.e., the properties of the distribution of change scores); (d) use of base-rate information to determine whether a change or discrepancy score is common or uncommon; and (e) *impairment* inferred on the basis of validity studies that demonstrate that large and relatively rare change scores are statistically more common in patient groups with a known condition of interest than would be expected among the reference population.

Although serial assessments share much in common with single-point assessments, they also pose unique interpretative challenges because two or more sets of scores are involved. Under ideal test–retest conditions, a patient's retest performance should be the same as that observed at baseline, and any change or deviation from baseline would be clinically relevant. However, in the absence of perfect test stability and reliability, the clinician must deal with the residuals of these statistical properties, namely, bias and error.

Bias Bias represents a systematic change in performance. The most important source of systematic bias in clinical practice is the variable of interest, that is, the effect of disease progression over time,

the impact of a surgical or pharmacological intervention, or the effect of rehabilitation. However, second only to the variable of interest, the most common source of bias in serial cognitive assessment is a positive *practice effect* in which performance is enhanced by previous test exposure, although negative biases can also occur such as those seen in aging [18]. For example, in a meta-analysis on practice effects on commonly used neuropsychological tests, Calamia et al. (2012) reported a mean practice effect of approximately +0.24 standard deviation units but noted that age decreased practice effects by approximately 0.004 per year after the age of 40 [19]. Other forms of systematic bias on retest performance are education, gender, clinical condition, baseline level of performance, and retest interval [19–22]. Where large, positive practice effects are expected, the absence of change may actually reflect a decrement in performance. To make accurate diagnoses, the clinician must separate the effects of the variable of interest from other sources of bias.

Error In addition to systematic biases, tests themselves are imperfect tools and can introduce an element of random error. For our purposes here, we will only consider two sources of error affecting serial assessment, both of which are inversely related to the test's reliability. The first is *measurement error* or the fidelity of the test, and it refers to the theoretical distribution of random variations in observed test scores around an individual's true score, which is characterized by the *standard error of measurement (SEM)*. Because the *SEM* is inversely related to a test's reliability, tests with low reliability (<0.70) have large *SEMs* surrounding a person's true score at both baseline and on retest, and large test–retest differences can occur simply as random fluctuations in measurement. Conversely, small test–retest changes can be reliable and clinically meaningful for tests with high reliability (>0.90). Test–retest reliabilities of 0.70 or greater are often considered to be the minimum acceptable standard for psychological tests in outcome studies [23], and practitioners should be wary when interpreting cognitive change scores on tests that have lower reliabilities.

The second source of error affecting change scores is *regression to the mean*, which refers to the susceptibility of retest scores to regress toward the mean of the scores at baseline. The more a score deviates from the population mean at baseline, the more likely it will regress back toward the mean on retest. How much a score regresses depends on the reliability of the test. Again, scores on tests with high reliability show less susceptibility to regression to the mean than those on tests with lower reliability. The bottom line for clinicians when planning to perform serial assessments and faced with two tests purported to assess the same cognitive construct—choose the one with the better reliability!

Alternate forms Alternate forms are often touted as an effective means for avoiding or minimizing practice effects due to test familiarity. Carefully constructed alternative forms may attenuate the effects of content-specific practice for some measures [24]. However, research demonstrates that alternate forms used in serial assessments still show significant practice effects [25]. While alternate forms may dampen practice effects due to content familiarity, they do not control for procedural learning and other factors that contribute to the overall practice effect. More importantly, rote use of alternate forms in serial assessment ignores other factors that impact interpretation of test–retest change scores, namely, reliability and error [17].

Reliable Change in Serial Assessments with Older Adults

It should be clear that the interpretation of test–retest change scores is not a straightforward matter, and making accurate diagnostic judgments about whether an older adult has shown significant deterioration (or improvement) in cognitive status over a retest interval requires us to consider the role of bias and error in our measurements. Bias and error are problems only to the degree that they are unknowns and not taken into account when interpreting change scores. In this section, we will discuss *reliable change methods*, a family of related

statistical procedures that attempt to take into account the impact of differential practice effects and other systematic biases, measurement error, and regression to the mean on the interpretation of change scores. We do not intend to do a comprehensive or in-depth review of these procedures, and the interested reader is directed to other sources for more complete coverage [15, 17, 21, 22, 26–28]. Rather, we wish to distil the essential features of reliable change methods and demonstrate how these tools can be used diagnostically to evaluate meaningful cognitive change in older adults.

Reliable Change: A Statistical Approach to Meaningful Change

To understand the concept of reliable change, we need to distinguish between what is statistically significant at a group level and what is clinically meaningful at the individual level. Repeated measure tests of statistical significance tell us whether the mean difference between two groups of a given magnitude is a reliable difference that would not be expected to occur by chance at some predefined probability level (e.g., $p < 0.05$). However, the base rates of such differences at the level of the individual may actually occur with some regularity even when no real behavioral difference. For this reason, Matarazzo and Herman have urged clinicians to routinely consider base-rate data in their clinical interpretation of test–retest evaluations [29].

Reliable Change: The Basic Model

Reliable change methods all fundamentally strive to evaluate the base rates of difference scores in a population and to determine whether the difference between scores for an individual is statistically rare and cannot be accounted for by various sources of bias (e.g., practice) or error (e.g., measurement error and regression to the mean). Like a ruler or yardstick that measures *change* from point A to point B along a standard metric (inches/yards), the basic form for any reliable change method is a ratio: reliable change

(RC) = (*change score*) / (*standard error*), where the *standard error* describes the dispersion of change scores that would be expected if no actual change had occurred [30]. This is simply the distribution of test–retest scores one would see in a reference population. RC is typically expressed as a standardized *z*-score under the unit curve that has a mean of 0 and a standard deviation of 1.0. The base rate of a given RC value being equal to the percentile associated with the *z*-score, for example, a *z*-score or RC of -1.64 , falls at the bottom fifth percentile. The various reliable change methods reported in the literature primarily vary along two dimensions: whether the *change score* in the numerator is a simple-difference or a predicted-difference score and whether the *standard error* in the denominator represents a measure of dispersion (observed or estimated) around the mean of difference scores or around a regression line.

Simple versus predicted-difference change scores For the *change score* component of the RC ratio, when we do follow-up evaluations on a patient, we generally look at the retest scores and compare them with the baseline score (retest–baseline) to see if the difference is positive or negative. This is the simple-difference approach. When no difference is expected over the retest interval (perfect stability), the simple-difference change score reflects the patient’s individual deviation from a population mean difference score of 0 or no expected change. However, as we have noted earlier, there are many sources of bias affecting retest scores, with practice often exerting a strong positive bias. As a result, the actual population mean of the test–retest *change scores* is positive and has led to the development of a practice-adjusted simple-difference approach [31]. For example, the mean retest performance on the Wechsler Memory Scale-III (WMS-III) Immediate Memory Index is 13.4 points higher than at baseline when readministered several weeks later [32]. If our 68-year-old male patient that we are following for suspected dementia has a baseline score of 97 and a retest score of 100, has he actually shown an improvement of 3 points when the average retest *change score* is 13.4 or a

decrement of -10.4 points ($13.4 - 3 = -10.4$) from expected change? To adjust for expected practice effects, Chelune and colleagues have suggested centering the *change score* component of the RC deviations around the mean of the expected practice effect and calculating the *change score* discrepancy from this mean [31].

The second approach to calculating the *change score* component of the RC ratio is the predicted-difference method. This is a regression-based approach that uses a patient's baseline performance to predict what his/or her retest score is expected to be at retest, with the regression equation being one derived from an appropriate reference sample. The discrepancy between the patient's actual observed retest score and the predicted retest score ($Y - Y'$) constitutes the *change score* discrepancy. Entering the baseline score as a predictor of the retest score into the regression equation allows practice effects to be modeled as a function of baseline performance (rather than as a constant) while also accounting for regression to the mean [33], two aspects not accounted for by the simple-difference approach. As in any regression approach, the equation can be univariate, using only the baseline score as the sole predictor, or multivariate, using additional information from other potential sources of bias as predictors, such as age, education, gender, and retest interval. In the example above of the 68-year-old male patient suspected of dementia, a regression-based equation using baseline WMS-III Immediate Memory Index scores and age was computed for the WMS-III test-retest standardization sample [17]. Given a baseline score of 97 for a 68-year-old normal individual, the predicted retest score would be 108.8. Our patient's predicted *change score* deviation is -8.8 points (observed retest score of 100 minus the predicted test score of 108.8). The reader will note that the -8.8 -point predicted *change score* discrepancy is smaller than the -10.4 -point simple-difference *change score*. The reason for this is that the regression-based predicted *change score* modeled not only practice effects (a positive bias) but also age (a negative bias), which dampened the expected practice effect, resulting in a smaller (although perhaps more accurate) expected retest score.

Measures of dispersion for the simple-difference method Once the individual's *change score* discrepancy has been computed, we have a measure of *change* but do not know whether the change is large or small without having a standard metric to evaluate the dispersion of *change scores* that would occur in the absence of real change (i.e., changes simply due to error). This is reflected in the denominator of the RC ratio, and the choice of the measure of dispersion has been the subject of much debate and refinement in the reliable change literature [15, 17, 22, 26, 27, 34]. The simplest version of the *standard error* component of the RC ratio is simply the *standard deviation* of the observed *change score* discrepancies. In our dementia case example with the WMS-III, the mean test-retest *change score* obtained from the WAIS-III/WMS-III Technical Manual is 13.4 [32]. However, like many test manuals and normative studies that report the means and standard deviations of the test and retest scores, the *standard deviation* of difference (change) scores was not reported. With permission from the test publisher, Chelune calculated the actual *standard deviation* of *change scores* for the WMS-III Immediate Memory Index from the retest sample and found it to be 10.2 [17]. With this measure of dispersion, we can calculate the RC magnitude of our patient's change score by dividing the observed practice-adjusted simple-difference score (-10.4) by the standard deviation of differences (10.2) and obtain an RC z -score of -1.02 . A z -score of this magnitude would be expected to occur in only about 15% of cases when no real change has occurred. Is this sufficiently rare to classify our patient's *change score* as meaningful? Most studies of reliable change invoke a 90% RC confidence interval (z -score ± 1.64), in which only 5% of cases would be above or below this level of change. For our patient's change score to reach this level of decline, he would have needed a retest score between 93 and 94. It is worth emphasizing that a seemingly minor decrement in performance (e.g., 3–4 standard score points in this case), a change that many clinicians might call "within the range of the test's variability," actually reflects a reliable change when corrected for expected practice effects and measurement error.

In the absence of having the actual *standard deviation* of difference scores, it is possible to estimate it in one of several ways. Jacobsen and Truax initially introduced the Reliable Change Index (RCI) as a means for calculating RC with only knowledge of the simple-difference *change score* and the *standard error of the difference scores* (S_{diff}), a measure of dispersion derived from *SEM* for the test at baseline [35]. Chelune and colleagues later adapted the RCI by adjusting for the mean practice effect [31]. In a further refinement, Iverson suggested a modified RCI that used the *SEM* at both baseline and at retest to calculate the S_{diff} [36]. Comparison of the two versions of the S_{diff} suggests that Iverson's method produces a closer estimate of the actual dispersion of *change scores* than that of Jacobsen and Truax. In the case of our WMS-III Immediate Memory example, the Iverson method produces a S_{diff} of 9.9 compared to 8.8 for the Jacobson and Truax method, where the actual *standard deviation* of differences was 10.2. A final common estimate of the observed dispersion of *change scores* is the *standard error of prediction*, which represents the standard error of a retest score predicted from a baseline score in a regression equation where the test reliability coefficient is the standardized beta coefficient [17]. In our WMS-III example, the *standard error of prediction* for the Immediate Memory Index is 10.1, very close to the observed standard deviation of actual change scores, namely, 10.2.

Standardized regression-based (SRB) approach. As noted in our discussion of the simple versus predicted methods of calculating the *change score* discrepancy in the RC ratio, the predicted-difference method generates predicted retest scores (Y') for individuals based on their specific baseline performances (X) using linear regression and then subtracts this from their observed retest scores (Y) to obtain their personal *change score* discrepancy ($Y - Y'$). Additional sources of potential bias (e.g., age, education, gender) can be added to the regression equation in a multivariate manner [33]. As noted earlier, this approach allows practice effects to be modeled as a function of individual baseline performance as well as accounting for regression

to the mean. This might be particularly important as these two variables interact (e.g., the practice effects may be attenuated by regression to the mean for someone with a high baseline score, whereas practice effects are enhanced by regression to the mean for an individual with a low initial baseline score). However, unlike the simple-difference approach where the standard error term in the denominator of the RC ratio reflects the dispersion of change scores around the mean of the *change scores*, the predicted-difference approach typically uses the *standard error of the estimate (SEE)* for the regression equation in the denominator of the RC ratio to reflect the dispersion of scores around the regression line. In our case example with the WMS-III Immediate Memory Index [17], the regression equation for predicting retest scores was given as:

$$Y' = (\text{Baseline score} * 1.00) + (\text{Age} * -0.097) + 18.45, \text{ with an } SEE \text{ of } 10.24$$

The first part of this equation gives us an individual's predicted retest score that can be used to calculate the *change-score* discrepancy component of the RC ratio, whereas the *SEE* gives us the standard error term for the denominator. The reader will note that the *SEE* for the regression line is the same as the observed *standard deviation* of the simple-difference *change scores*.

While several authors have noted that the various RC methods produce relatively similar results [22, 30], the SRB RC-approach has generally become the preferred method for individual prediction, provided that the clinician has access to prediction equations derived from reference samples appropriate to their patients. While there is a growing body of such SRB equations for a variety of tests commonly used with older adults [8, 9, 16, 20, 37, 38], and some tests such as the fourth edition of the Wechsler Adult Intelligence and Memory Scales have incorporated RC algorithms into their scoring software [10], there is still a paucity of published longitudinal SRB data. Fortunately, as will be seen in the next section, John Crawford and Paul Garthwaite have developed a simple but powerful tool for building

regression equations from summary data that can be applied to the individual case [39].

Regression models of reliable change derived from summary data As noted by Crawford and Garthwaite [39], not all neuropsychologists are aware that it is possible to construct regression equations for predicting an individual's retest performance from their baseline performance simply using sample summary data, for which there is a potential wealth of clinically useful information available in test manuals and the published literature. To build univariate regression equations from summary data alone, one only needs the means and standard deviations for test and retest scores, the size of the sample, and the test–retest reliability coefficient (or alternately the *t*-value from a pair-samples *t* test). In their 2007 paper, Crawford and Garthwaite delineate the statistical steps necessary to build such regression equations, as well as the further steps needed to compute the associated statistics for drawing inferences concerning the individual case. Recognizing that the computations involved are tedious and prone to error, Crawford and Garthwaite also developed a compiled calculator

that is available for download at no cost from the following web address: <http://www.abdn.ac.uk/~psy086/dept/regbuild.htm>

To use this calculator, one only need input the sample summary data and the patient-specific test–retest scores. Using the summary data from Chelune [17], Table 5.1 illustrates the output generated for our hypothetical 68-year-old patient whose baseline Immediate Memory Index was 97 at baseline and 100 on retest. The output is remarkably similar to that presented in previous sections for our patient example using various RC methods. Generally, the various approaches would predict our patient to have a retest score of 109–110 given his baseline score of 97. His observed retest score of 100 is 9–10 points below expectations (RC *z*-score deviation of about -1.0), which would likely occur in only about 15% of a sample for which there were no significant intervening events affecting cognition.

Although the Crawford and Garthwaite's regression calculator presented here is univariate [39], it has recently been expanded to handle multiple predictors, and this executable calculator is also available for download online at

Table 5.1 Output from Crawford and Garthwaite's [39] calculator to build regression equations from sample summary data for a hypothetical patient with test–retest scores of 97 and 100 on the Wechsler Memory Scale-III Immediate Memory Index

<i>Inputs</i>	
Mean for predictor variable (<i>X</i>) in sample used to build the equation =	100.2
Standard deviation for predictor variable (<i>X</i>) in sample =	15.9
Mean for the criterion variable (<i>Y</i>) in sample =	113.7
Standard deviation for the criterion variable (<i>Y</i>) in sample =	19.2
Correlation between predictor and criterion variable =	0.85
Sample size =	297
Individual's score on the predictor (<i>X</i>) variable =	97
Individual's obtained score on <i>Y</i> =	100
<i>Outputs</i>	
Regression equation built from the summary data: $Y = 10.8532 + (1.0264 * X)$	
Standard error of estimate for regression equation =	10.1314
<i>Analysis of the individual case</i>	
Individual's predicted score from regression equation =	110.4155
Discrepancy (obtained minus predicted) between individual's obtained and predicted scores =	-10.4155
Standardized discrepancy between individual's obtained and predicted scores =	-1.0262
Significance test (<i>t</i>) on the standardized discrepancy between individual's obtained and predicted scores:	
One-tailed probability =	0.1528
Estimated percentage of population obtaining a discrepancy more extreme than individual =	15.280799%

http://www.abdn.ac.uk/~psy086/dept/RegBuild_MR.htm [40].

Advanced concepts and models of reliable change The various RC methods we have described so far only consider measuring change as a discrete event across two points in time. However, there are many clinical situations where individuals are assessed serially across multiple time points, and change may be better described in terms of *trajectories of change* and intraindividual *rates of cognitive decline*. Early attempts to assess reliable change across multiple time points either averaged reliability coefficients and measures of dispersion between the various time points to arrive at composite indices of RC [41] or computed separate RC indices between each pair of time points [38]. Recently, more innovative approaches have been employed to model *change* as a trajectory or slope across multiple time points.

It is beyond the scope of this chapter to do more than alert the reader to some of these innovative approaches and to provide exemplars. Some investigators are using regression models that attempt to predict an individual's performance at time point $t_2 + n$ by entering into regression formula not only baseline performance but the practice effects between previous time points. For example, Duff and associates [8] developed multivariate SRB equations for several neuropsychological tests widely used with older adults that used baseline performance, demographic variables, and short-term practice effects (baseline to 1 week) in predicting retest scores 1 year later. Attix and colleagues [42] developed SRB normative neuropsychological trajectories for a variety of test measures administered five times at 6-month intervals by entering in successive performances at each time point as predictors of subsequent performance at the next time point. Other investigators have focused on developing regression models that compare an individual's slope of performance across multiple time points to that of a control sample [43, 44]. Still others are using variations of longitudinal linear mixed models to estimate age-adjusted mean slopes and

confidence intervals of change to identify individuals whose performances begin to deviate from expectation [7, 45]. Growth mixture modeling has also been applied to longitudinal data sets to identify subgroups of individuals who show different cognitive trajectories over time [46–49]. Clearly, we are on the verge of seeing a new generation of RC methods to assess reliable change in patients' performances over time.

A Case Example: Application of Reliable Change Methods in Clinical Practice

The accumulation of pathophysiological changes characteristic of Alzheimer's disease (AD) is believed to develop years, if not decades, before the clinical expression of frank memory loss and general cognitive decline [50]. To maximize the efficacy of emerging disease-modifying therapies and to support continued functional independence, early detection of Alzheimer's disease (AD) and other neurodegenerative disorders is paramount [46, 51]. Descriptive clinical states such as *cognitive impairment but not dementia* (CIND) and MCI have been introduced to describe abnormal cognitive states that place individuals at increased risk for progressing to AD [52]. However, these clinical states describe individuals who are already symptomatic. One does not wake up one day with dementia or MCI. Rather, cognitive decline, like neurodegenerative disease, is a dynamic process that evolves over time. Hence, serial neuropsychological evaluations have come to play an important role in documenting cognitive decline in geriatric settings.

Let us consider a case example of a 63-year-old, right-handed man with a Ph.D. Our patient is a successful professor of sociology at a major university and a married father of three children. His past medical history is significant for depression and some cardiac issues, both currently well controlled. He has been stable on his medications for many years, and they are not thought to be an issue with respect to cognition. Our patient has noticed insidious and progressive memory difficulties for about 2 years and presents to our cognitive disorders

clinic for evaluation. His neurologist obtains a Mini-Mental State Exam score of 30/30 but on further bedside testing notes some subtle memory difficulties. The neurologist decides to refer the patient to us for comprehensive neuropsychological evaluation. We perform our evaluation and find that the patient has a relatively circumscribed pattern of memory deficit within the context of otherwise normal findings (see baseline scores in Table 5.2). Our impression is that this patient has amnesic MCI. We know from the research literature that patients with MCI have an increased risk of showing further decline and developing a frank dementia. However, we also know that some of these individuals revert back to “normal” when seen in follow-up [53, 54]. We share these observations with our referring neurologist and recommend that the patient be referred for a follow-up evaluation in 1 year to assess whether there has been any evidence of significant interval change in his neurocognitive status. Seeing the wisdom in our recommendations, the neurologist agrees and orders repeat testing in a year.

The patient returns 12 months later, and we repeat his evaluation. As we can see from the test–retest data summarized in Table 5.2, some of our patient’s scores have gotten worse and some have gotten better. To understand which of these changes are reliable and meaningful given the different psychometric properties of the tests in our battery and to place them on a common metric, we turned to RC methods. For our purposes here, we computed reliable change information using the predicted-difference method. Using the test–retest data presented in the manuals for the tests or from longitudinal research studies with samples of healthy older adults, we entered the sample summary data into Crawford and Garthwaite’s regression calculator [39] along with our patient’s baseline and retest scores. In the right-hand columns of Table 5.2, we present the patient’s predicted retest scores given his baseline performances, the observed–predicted discrepancy ($Y - Y'$), and the associated z -scores and population percentiles associated with the predicted-difference discrepancies. From these data, we can see that the patient’s memory has continued to significantly deteriorate. We also

note that his global mental status on the Mattis Dementia Rating Scale [55] and on the WAIS-III verbal comprehension index [32] shows signs of notable deterioration. At this point, we can confidently say that the patient’s current test results reflect some further deterioration in his capacity to learn and remember new information as well as some increased difficulties with verbal intellectual abilities. While he is still likely to meet the criteria for MCI rather than dementia, his increased difficulties with verbal skills are worrisome for a neurodegenerative disorder such as Alzheimer’s disease.

Future Directions: Change as a Neurocognitive Biomarker

As noted earlier, practice effects are defined as improvements in test scores due to repeated exposure to the testing materials. Traditionally, practice effects have been viewed as error variance that need to be controlled, managed, or otherwise accounted for in our interpretation of change. However, practice effects, like cognitive change in general, seem to be a unique variable that can potentially provide clinically useful information about diagnosis, prognosis, presence of brain pathology, and treatment recommendations for our patients [59]. Over the past several years, we have been prospectively examining practice effects as a neurocognitive biomarker in the development of dementia in older adults.

In an initial study examining practice effects in community-dwelling seniors with MCI, we observed two subgroups: those that benefited from practice across 1 week and those that did not [60]. Those that showed significant gains after repeat testing could no longer be classified as MCI, as they now appeared intact. These MCI participants might reflect “accidental” MCI [53, 54]. Conversely, the MCI participants that did not benefit from practice retained their original diagnostic classification, and these participants more likely demonstrate the construct of MCI. In this way, short-term practice effects provide diagnostic information that was not available with baseline

Table 5.2 Clinical case example of test–retest scores and reliable change (RC) information based on data in bold using Crawford and Garthwaite’s [39] approach to derive RC regression equation from sample summary data

Test	Baseline scores			Follow-up scores			Reliable change (RC) information			
	Raw	Standard score	T-score	Raw	Standard score	T-score	Predicted retest score	Discrepancy (Y – Y')	RC z-score	Population percentile
<i>Global mental status</i>										
Mini-mental state exam ^a	29			28			28.57	–0.58	–0.33	37
Mattis Dementia Rating Scale: Total ^b	140	11		133	8		135.5	–6.36	–1.48	7
<i>Wechsler tests</i>										
Test of premorbid functioning ^c		125			118		124.25	–6.25	–1.13	13
<i>WAIS-III Adult Intelligence Scale^d</i>										
General ability index	82	126	58	77	119	43				
Verbal comprehension	46	131	61	40	118	49	130.43	–12.43	–3.07	<1
Perceptual organization	36	111	47	37	114	50	115.11	–1.11	–0.19	43
Processing speed	19	96	41	20	99	43	100.14	–1.14	–0.17	43
<i>Memory measures</i>										
WMS-III memory ^e										
Logical memory-immediate	30	8	35	20	4	21	9.88	–5.88	–3.14	<1
Logical memory-delayed	14	8	37	4	3	18	10.61	–7.61	–3.87	<1
Digit span ^d	20	13	53	21	13	53	13.26	–0.26	–0.20	42
<i>Hopkins Verbal Learning Test^f</i>										
Total trials 1–3	22		37	17		28	23.36	–6.36	–1.35	9
Delay	0		<20	0		<20	2.61	–2.61	–0.95	17
<i>Brief Visuospatial Memory Test^f</i>										
Trials 1–3	12		31	2		<20	13.46	–11.46	–2.21	<1
Delay	0		<20	0		<20	2.25	–2.25	–1.01	16
<i>Language</i>										
Boston Naming Test ^g	58	13		58	13		57.72	0.28	0.12	55
Controlled Oral Word Association ^f	46	13		49	13		45.32	3.68	0.41	66
<i>Visuospatial functions</i>										
Judgment of line orientation ^h	30	16		28	14		23.48	4.52	0.83	79
KBNA complex figure and clock drawing total ⁱ	54	12		55	14		10.98	3.02	1.21	88
<i>Executive functions</i>										
Trail-making A time ^f	38	8		33	10		39.39	–6.93	0.44	67
Trail-making B time ^f	63	11		97	8		70.89	26.12	–0.51	30
KBNA practical problem and conceptual shifting total ⁱ	29	13		29	13		11.53	1.47	0.58	70

Notes: Sources of normative data used in developing RC prediction equations

^aTombaugh [38]

^bPedraza et al. [55]

^cHoldnack and Drozdick [10]

^dThe Psychological Corporation [32]; Table 3.8

^eThe Psychological Corporation [32]; Table 3.11

^fDuff et al. [8]

^gDuff et al. [56]

^hDuff et al. [57]

ⁱDarby et al. [58]

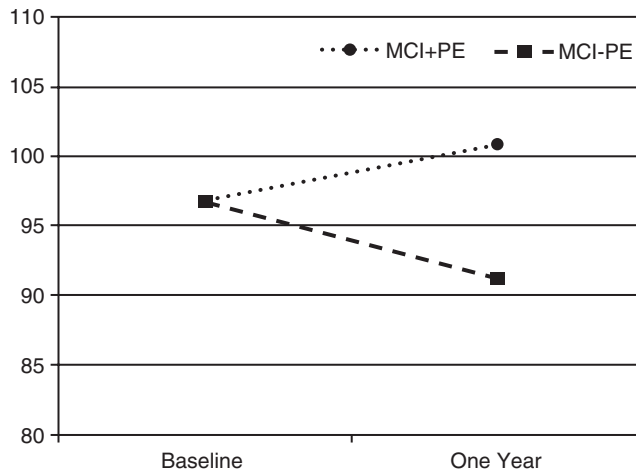


Fig. 5.1 Cognitive change across 1 year in patients with differential practice effects. Note MCI+PE=individuals with mild cognitive impairment who showed large practice effects across 1 week; MCI-PE=individuals with mild

cognitive impairment who showed minimal practice effects across 1 week; y-axis=age-corrected standard score ($M=100$, $SD=15$) on total scale score of the Repeatable Battery for the Assessment of Neuropsychological Status

data. Others also have found practice effects to be diagnostically useful in MCI [61].

Prognostically, the presence of practice effects suggests a better outcome, whereas the absence of practice effects suggests a poorer outcome. In two independent samples of individuals with MCI, we have observed that practice effects predict future cognition, above and beyond baseline cognition [8, 62]. As seen in Fig. 5.1, when we followed our two MCI subgroups across 1 year, those that benefitted from practice across 1 week tended to remain cognitively stable across 1 year, and those that did not show the expected practice effects across 1 week tended to decline across 1 year [63].

In a sample of 25 older adults without dementia (some intact, some with MCI), we observed that practice effects across 1 week were negatively associated with amyloid deposition using F-18 flutemetamol positron-emission tomography (PET) imaging [64]. As seen in Fig. 5.2, smaller than expected practice effects (i.e., lower values on the x -axis) were seen in subjects with greater amyloid deposition (i.e., greater values on the y -axis). In this same cohort, we also noted that smaller practice effects across 1 week were associated with brain metabolism on fluorodeox-

ylucose (FDG) PET imaging, such that smaller practice effects were associated with brain hypometabolism [65].

Lastly, we have examined the utility of practice effects in predicting treatment response. In a small sample of community-dwelling and cognitively intact older adults, within-session practice effects predicted response to a memory training course: those that showed practice effects displayed larger gains related to the cognitive intervention than those that did not show robust practice effects [66]. Although these findings need to be replicated, practice effects appear to contribute to a clinician's decision about diagnosis, prognosis, brain pathology, and treatment response, especially in older adults with memory difficulties.

Conclusion

The assessment of cognitive change lies at the very heart of clinical neuropsychology. Understanding change and how we assess it with our various test measures is complex and challenging, yet given an appropriate conceptual framework and some simple statistical tools, it is something that neuropsychologists can do uniquely well. Test-retest practice effects

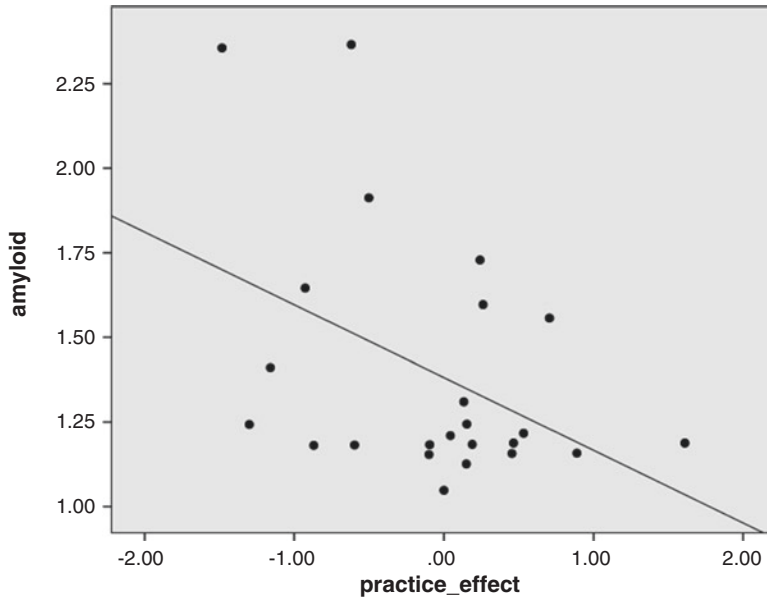


Fig. 5.2 Practice effect across 1 week is associated with amyloid deposition in non-demented older adults. On the x-axis, lower values reflect smaller than expected practice effects. On the y-axis, greater values reflect more amyloid deposits

are not simply statistical artifacts and something to be suppressed but rather something to be understood. Especially among older adults, the capacity to learn and benefit from exposures to new experiences to potentially guide future behavior has adaptive value and may be a biological marker of neural integrity that has diagnostic significance.

Clinical Pearls

- Patients deserve empirically based clinical decisions and recommendations.
- Test–retest change scores are unique variables with their own statistical and clinical properties that are different from the test measures from which they were derived.
- Where large positive practice effects are expected, the absence of change may actually reflect a decrement in performance.
- When planning to perform serial assessments and faced with two tests purported to assess the same cognitive construct, choose the one with the better reliability.
- Use of alternate forms in serial assessment may attenuate, but not eliminate, practice effects and do not address other factors that affect the interpretation of change scores, namely, bias and error.
- Test–retest reliabilities of 0.70 or greater are often considered to be the minimum acceptable standard for psychological tests in outcome studies, and practitioners should be wary when interpreting cognitive change scores on tests that have lower reliabilities.
- The basic form for any reliable change method is a ratio: $\text{reliable change (RC)} = (\text{change score}) / (\text{standard error})$, where the standard error describes the dispersion of change scores that would be expected if no actual change had occurred.
- The various reliable change methods reported in the literature primarily vary along two dimensions: (a) whether the *change score* in the numerator is a simple-difference or a predicted-difference score and (b) whether the *standard error* in the denominator represents a measure of dispersion (observed or estimated)

around the mean of difference scores or around a regression line.

- Not all neuropsychologists are aware that it is possible to construct regression equations for predicting an individual's retest performance from his/her baseline performance by simply using sample summary data, for which there is a potential wealth of clinically useful information available in test manuals and the published literature.
- For computing regression equations using sample summary data for individual cases, see Crawford and Garthwaite's univariate online calculator, and enter your patient-specific test-retest scores: <http://www.abdn.ac.uk/~psy086/dept/regbuild.htm>. For multivariate data, see the website at http://www.abdn.ac.uk/~psy086/dept/RegBuild_MR.htm.
- Although traditionally viewed as a source of bias, practice effects may provide valuable information about a patient's diagnosis, prognosis, brain pathology, and treatment response, especially for older adults with memory difficulties.

References

1. Barth JT, Pliskin N, Axelrod B, et al. Introduction to the NAN 2001 definition of a clinical neuropsychologist. NAN policy and planning committee. *Arch Clin Neuropsychol*. 2003;18:551–5.
2. Winblad B, Palmer K, Kivipelto M, et al. Mild cognitive impairment—beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *J Intern Med*. 2004;256:240–6.
3. Salthouse TA. Does the meaning of neurocognitive change change with age? *Neuropsychology*. 2010;24:273–8.
4. Howieson D, Holm LA, Kaye JA, Oken BS, Howieson J. Neurologic function in the optimally healthy oldest old: neuropsychological evaluation. *Neurology*. 1993;43:1882–6.
5. Cullum CM, Butters N, Troster AI, Salmon DP. Normal aging and forgetting rates on the Wechsler memory scale—revised. *Arch Clin Neuropsychol*. 1990;5:23–30.
6. Busch RM, Chelune GJ, Suchy Y. Using norms in the neuropsychological assessment of the elderly. In: Attix DK, Welsh-Bohmer KA, editors. *Geriatric neuropsychology: assessment and intervention*. New York: Guilford; 2006. p. 133–57.
7. Howieson D, Carlson NE, Moore MM, et al. Trajectory of mild cognitive impairment onset. *J Int Neuropsychol Soc*. 2008;14:192–8.
8. Duff K, Beglinger LJ, Moser DJ, Paulsen JS, Schultz SK, Arndt S. Predicting cognitive change in older adults: the relative contribution of practice effects. *Arch Clin Neuropsychol*. 2010;25:81–8.
9. Duff K. Predicting premorbid memory functioning in older adults. *Appl Neuropsychol*. 2010;17:278–82.
10. Holdnack JA, Drozdick LW. *Advanced clinical solutions for WAIS-IV and WMS-IV clinical and interpretative manual*. San Antonio: Pearson; 2009.
11. Chelune GJ. Evidence-Based Practices in Neuropsychology. In: Bowden S, editor. *Neuropsychological assessment in the age of evidence-based practice*. New York: Oxford; 2017.
12. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992;268(17):2420–5.
13. Chelune GJ. Evidence-based research and practice in clinical neuropsychology. *Clin Neuropsychol*. 2010;24:454–67.
14. Heilbronner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, Hart RP. Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: the utility and challenges of repeat test administrations in clinical and forensic contexts. *Clin Neuropsychol*. 2010;24:1267–78.
15. Hinton-Bayre AD. Deriving reliable change statistics from test-retest normative data: comparison of models and mathematical expressions. *Arch Clin Neuropsychol*. 2010;25:244–56.
16. Stein J, Luppia M, Braehler E, Konig HH, Riedel-Heller SG. The assessment of changes in cognitive functioning: reliable change indices for neuropsychological instruments in the elderly—a systematic review. *Dement Geriatr Cogn Disord*. 2010;29:275–86.
17. Chelune GJ. Assessing reliable neuropsychological change. In: Franklin R, editor. *Prediction in forensic and neuropsychology: new approaches to psychometrically sound assessment*. Mahwah: Erlbaum; 2003.
18. Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*. 2010;24:563–72.
19. Calamia M, Markon K, Tranel D. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin Neuropsychol*. 2012;26(4):543–70.
20. Duff K, Schoenberg MR, Patton D, et al. Regression-based formulas for predicting change in RBANS subtests with older adults. *Arch Clin Neuropsychol*. 2005;20:281–90.
21. Lineweaver TT, Chelune GJ. Use of the WAIS-III and WMS-III in the context of serial assessments:

- interpreting reliable and meaningful change. In: Tulskey DS, Saklofske DH, Chelune GJ, et al., editors. *Clinical interpretation of the WAIS-III and WMS-III*. New York: Academic; 2003. p. 303–37.
22. Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: a comparison of four models. *J Int Neuropsychol Soc*. 1999;5:357–69.
 23. Slick DJ. Psychometrics in neuropsychological assessment. In: Strauss E, Sherman EMS, Spreen O, editors. *A compendium of neuropsychological tests*. New York: Oxford University Press; 2006. p. 1–43.
 24. Benedict RHB, Zgaljardic DJ. Practice effects during repeated administrations of memory tests with and without alternate forms. *J Clin Exp Neuropsychol*. 1998;20:339–52.
 25. Beglinger LJ, Gaydos B, Tangphao-Daniels O, et al. Practice effects and the use of alternate forms in serial neuropsychological testing. *Arch Clin Neuropsychol*. 2005;20:517–29.
 26. Collie A, Darby DG, Falletti MG, Silbert BS, Maruff P. Determining the extent of cognitive change after coronary surgery: a review of statistical procedures. *Ann Thorac Surg*. 2002;73:2005–11.
 27. Crawford JR. Quantitative aspects of neuropsychological assessment. In: Goldstein LH, McNeil JE, editors. *Clinical neuropsychology: a practical guide to assessment and management for clinicians*. 2nd ed. Chichester: Wiley; 2011.
 28. Duff K. Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. *Arch Clin Neuropsychol*. 2012;27(3):248–61.
 29. Matarazzo JD, Herman DO. Base rate data for the WAIS-R: test-retest stability and VIQ-PIQ differences. *J Clin Neuropsychol*. 1984;6:351–66.
 30. Hinton-Bayre AD. Methodology is more important than statistics when determining reliable change. *J Int Neuropsychol Soc*. 2005;11:788–9.
 31. Chelune GJ, Naugle RI, Luders H, Sedlak J, Awad IA. Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology*. 1993;7:41–52.
 32. The Psychological Corporation. *Updated WAIS-III WMS-III technical manual*. San Antonio: The Psychological Corporation; 1997.
 33. McSweeny AJ, Naugle RI, Chelune GJ, Luders H. “T-scores for change”: an illustration of a regression approach to depicting change in clinical neuropsychology. *Clin Neuropsychol*. 1993;7:300–12.
 34. Maassen GH. Principles of defining reliable change indices. *J Clin Exp Neuropsychol*. 2000;22:622–32.
 35. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991;59:12–9.
 36. Iverson GL. Interpreting change on the WAIS-III/WMS-III in clinical samples. *Arch Clin Neuropsychol*. 2001;16:183–91.
 37. Knight RG, McMahon J, Skeaff CM, Green TJ. Reliable change index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Arch Clin Neuropsychol*. 2007;22:513–8.
 38. Tombaugh TN. Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Arch Clin Neuropsychol*. 2005;20:485–503.
 39. Crawford JR, Garthwaite PH. Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*. 2007;21:611–20.
 40. Crawford JR, Garthwaite PH, Denham AK, Chelune GJ. Using regression equations built from summary data in the psychological assessment of the individual case: extension to multiple regression. *Psychol Assess*. 2012;24:801.
 41. Ivnik RJ, Smith GE, Lucas JA, et al. Testing normal older people three or four times at 1- to 2-year intervals: defining normal variance. *Neuropsychology*. 1999;13:121–7.
 42. Attix DK, Story TJ, Chelune GJ, et al. The prediction of change: normative neuropsychological trajectories. *Clin Neuropsychol*. 2009;23:21–38.
 43. Chelune GJ, Ivnik R, Smith G. Application of reliable change methods for identifying abnormal rates of cognitive decline in dementia. *Alzheimers Dement*. 2006;2(S1):374.
 44. Crawford JR, Garthwaite PH. Statistical methods for single-case studies in neuropsychology: comparing the slope of a patient’s regression line with those of a control sample. *Cortex*. 2004;40:533–48.
 45. Darby DG, Pietrzak RH, Fredrickson J, Woodward M, Moore L, Fredrickson A, Sach J, Maruff P. Intra-individual cognitive decline using a brief computerized cognitive screening test. *Alzheimers Dement*. 2012;8(2):95–104.
 46. Albert M, Blacker D, Moss MB, Tanzi R, McArdle JJ. Longitudinal change in cognitive performance among individuals with mild cognitive impairment. *Neuropsychology*. 2007;21:158–69.
 47. Stern Y, Liu X, Albert M, et al. Application of a growth curve approach to modeling the progression of Alzheimer’s disease. *J Gerontol A Biol Sci Med Sci*. 1996;51:M179–84.
 48. Small BJ, Backman L. Longitudinal trajectories of cognitive change in preclinical Alzheimer’s disease: a growth mixture modeling analysis. *Cortex*. 2007;43:826–34.
 49. Hayden KM, Reed BR, Manly JJ, et al. Cognitive decline in the elderly: an analysis of population heterogeneity. *Age Aging*. 2011;40(6):684–9.
 50. Morris JC, Storandt M, Miller JP, et al. Mild cognitive impairment represents early-stage Alzheimer disease. *Arch Neurol*. 2001;58:397–405.
 51. Amieva H, Le Goff M, Millet X, et al. Prodromal Alzheimer’s disease: successive emergence of the clinical symptoms. *Ann Neurol*. 2008;64:492–8.
 52. Petersen RC, Smith GE, Waring SC, Ivnik RJ, Kokmen E, Tangelos EG. Aging, memory, and

- mild cognitive impairment. *Int Psychogeriatr*. 1997;9(Suppl 1):65–9.
53. Brooks BL, Iverson GL, White T. Substantial risk of “accidental MCI” in healthy older adults: base rates of low memory scores in neuropsychological assessment. *J Int Neuropsychol Soc*. 2007;13:490–500.
 54. de Rotrou J, Wenisch E, Chausson C, Dray F, Faucounau V, Rigaud AS. Accidental MCI in healthy subjects: a prospective longitudinal study. *Eur J Neurol*. 2005;12:879–85.
 55. Pedraza O, Smith GE, Ivnik RJ, et al. Reliable change on the dementia rating scale. *J Int Neuropsychol Soc*. 2007;13:716–20.
 56. Duff K, Callister C, Dennett K, Tometich D. Practice effects: a unique cognitive variable. *Clin Neuropsychol*. 2012;26(7):1117–27.
 57. Duff K, Beglinger LJ, Van Der Heiden S, et al. Short-term practice effects in amnesic mild cognitive impairment: implications for diagnosis and treatment. *Int Psychogeriatr*. 2008;20:986–99.
 58. Darby D, Maruff P, Collie A, McStephen M. Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology*. 2002;59:1042–6.
 59. Duff K, Beglinger LJ, Schultz SK, et al. Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch Clin Neuropsychol*. 2007;22:15–24.
 60. Duff K, Lyketsos CG, Beglinger LJ, et al. Practice effects predict cognitive outcome in amnesic mild cognitive impairment. *Am J Geriatr Psychiatry*. 2011;19(11):932–9.
 61. Duff K, Foster NL, Hoffman JM. Practice effects and amyloid deposition: preliminary data on a method for enriching samples in clinical trials. *Alzheimer Dis Assoc Disord*. 2014;28(3):247–52.
 62. Duff K, Horn KP, Foster NL, Hoffman JM. Short-term practice effects and brain hypometabolism: preliminary data from an FDG PET study. *Arch Clin Neuropsychol*. 2015;30(3):264–70.
 63. Duff K, Beglinger LJ, Moser DJ, Schultz SK, Paulsen JS. Practice effects and outcome of cognitive training: preliminary evidence from a memory training course. *Am J Geriatr Psychiatry*. 2010;18:91.
 64. Zec RF, Markwell SJ, Burkett NR, Larsen DL. A longitudinal study of confrontation naming in the “normal” elderly. *J Int Neuropsychol Soc*. 2005;11:716–26.
 65. Montse A, Pere V, Carme J, Francesc V, Eduardo T. Visuospatial deficits in Parkinson’s disease assessed by judgment of line orientation test: error analyses and practice effects. *J Clin Exp Neuropsychol*. 2001;23:592–8.
 66. Leach L, Kaplan E, Rewilak D, Richards B, Kaplan KBNA baycrest neurocognitive assessment: manual. San Antonio: The Psychological Corporation; 2000.