Current Topics in Science and Practice

Evidence-Based Indicators of Neuropsychological Change in the Individual Patient: Relevant Concepts and Methods

Kevin Duff*

Center for Alzheimer’s Care, Imaging, and Research, Department of Neurology, University of Utah, Salt Lake City, UT, USA

*Corresponding author at: Center for Alzheimer’s Care, Imaging and Research, Department of Neurology, University of Utah, 650 Komas Drive #106-A, Salt Lake City, UT 84108, USA. Tel.: +1-801-585-9983; fax: +1-801-581-2483.
E-mail address: kevin.duff@hsc.utah.edu (K. Duff).

Accepted 29 December 2011

Abstract

Repeated assessments are a relatively common occurrence in clinical neuropsychology. The current paper will review some of the relevant concepts (e.g., reliability, practice effects, alternate forms) and methods (e.g., reliable change index, standardized based regression) that are used in repeated neuropsychological evaluations. The focus will be on the understanding and application of these concepts and methods in the evaluation of the individual patient through examples. Finally, some future directions for assessing change will be described.

Keywords: Reliable change; Practice effects; Assessment

Introduction

Repeated assessments are a relatively common occurrence in clinical neuropsychology. Two or more testing sessions can be used to follow the natural progression of a condition, such as a dementia re-evaluation. Similarly, they can be used to track recovery after a neurological insult (e.g., improvements following traumatic brain injury or stroke). Serial cognitive evaluations may be used to evaluate the effectiveness of an intervention (e.g., temporal lobectomy, tumor resection, cognitive rehabilitation). The same individual might also be examined multiple times in the course of a forensic evaluation (e.g., seen by plaintiff and defense neuropsychologists). Although repeated neuropsychological assessments occur less frequently than single assessments, the former can be more complex than the latter. Since a recent policy paper by the American Academy of Clinical Neuropsychology (Heilbronner et al., 2010) recommended that neuropsychologists become more informed about the benefits and challenges associated with serial assessment, the current paper will review some of the relevant concepts and methods that are used in repeated neuropsychological evaluations. The focus of the paper will be on the understanding and application of these concepts and methods in the evaluation of the individual patient.

Concepts Associated with Change

In the classic test theory model, an observed score is some combination of a true score and error. Following this same logic, an observed change in test scores is likely some combination of true change and error. The true change is the proportion of variance with which neuropsychologists are most interested. If it could be isolated, this true change could reflect the actual disease progression, normal recovery from injury, or benefits of treatment. The error is the proportion of variance that could lead neuropsychologists astray in their interpretations and conclusions. As in a single assessment, error could reflect...
any systematic or random bias in the data, such as patient fatigue, poor lighting, or errors in test administration. In repeated assessments, these biases can be compounded with two or more testing sessions. For example, a patient may be equally fatigued at both assessments or more fatigued at one of the two assessments. Sources of error that are most relevant to repeated assessments can be grouped into three domains: variables associated with the test, variables associated with the testing situation, and variables associated with individual patient.

Variables Associated with the Test

Reliability. Typically defined as the degree to which a test score is systematic and free from error, reliability is often presented as a correlation, ranging from +1.0 (e.g., as $x$ increases, $y$ increases) to 0.0 (e.g., no relationship between $x$ and $y$) to −1.0 (e.g., as $x$ increases, $y$ decreases). However, a strong correlation does not necessarily imply that a test is good, yields stable scores, or accurately detects change. A strong correlation simply means that individuals retain their relative position within the distribution of scores from one testing session to the next. For example, the first two columns in Table 1 reflect Time 1 and Time 2 scores ($M = 100, SD = 15$) on the same test for a small sample. For these individuals, their scores on Time 2 are exactly the same as their scores on Time 1 (i.e., no change), which yields a correlation of +1.0. If these individuals displayed a slight improvement on time 1 (e.g., in third column, all scores increase by 1), then the correlation remains +1.0. If these individuals all dramatically drop (e.g., in the fourth column, all scores decrease by 40), the correlation is again +1.0. Regardless of the size of the change, if all individuals change by the same amount and retain their relative position within the group, the correlation does not change. In the fifth column, all individuals change slightly at Time 2 (e.g., some scores increasing and some decreasing by 1). This slight change dramatically alters the relative position in the distribution between Times 1 and 2, which leads to a correlation of .6. In the final column, small but inconsistent changes in the relative order of the individuals from Time 1 to Time 2 leads to a correlation of 0.0 (i.e., no relationship between Time 1 and Time 2 scores). In this example, reliability can be viewed as the degree to which individuals retain their relative position from Time 1 to Time 2. But, as will be discussed later, many factors can affect changes in ordering of individuals on retesting.

Even though reliability does not tell the whole story about assessing change, it is one of the key elements in nearly all statistical procedures for evaluating change. Therefore, several points should be mentioned. First, despite there being multiple types of reliability (e.g., internal consistency, inter-rater, parallel forms), test–retest reliability (or stability) is the most relevant in repeated assessments. Second, test–retest reliability is affected by the time interval between initial and repeated assessments. Shorter retest intervals lead to higher reliability coefficients, and longer retest intervals lead to lower reliability values. For example, on the Brief Visuospatial Memory Test-Revised, the manual (Benedict, 1997) reports a test–retest correlation of .86 across 55 days, whereas we have observed lower correlations ($r = .63$) on this same measure across 1 year (Duff, Beglinger, Moser, & Paulsen, 2010). Not surprisingly, most test manuals report test–retest correlations across relatively short intervals (e.g., days to weeks); intervals that are far shorter than most clinical retesting scenarios (e.g., months to years). Third, individual difference variables of the patient can affect reliability values. For example, the Wechsler Adult Intelligence Scale-IV (WAIS-IV) manual (Wechsler, 2008), younger adults tend to have higher test–retest correlations than older adults (Visual Puzzles: younger $r = .74$, older $r = .57$). Although there is little evidence in the literature, it is expected that other patient variables (e.g., education, intellect, diagnostic condition) could also affect reliability estimates. Lastly, not all cognitive domains yield the same reliability values. For example, in a large cohort of cognitively normal seniors tested on multiple occasions (Ivnik et al., 1999), higher retest correlations were observed for Verbal Comprehension ($r = .87$) and Attention-Concentration ($r = .81$) factors than for Learning ($r = .70$) and Retention ($r = .55$) factors. Not surprisingly, crystallized intelligence seems to be more stable than other cognitive processes. Finally, it should be noted that clinicians will have many options when seeking test–retest reliability coefficients for their individual patients. Nearly all test manuals report test–retest reliability data. Many journal articles with repeated testing will present some correlations. (Surprisingly, some published longitudinal studies, including some of our own, do not report this critical information, and we encourage authors of studies on

<table>
<thead>
<tr>
<th>Time 1</th>
<th>Time 2: no change</th>
<th>Time 2: small change, retain position</th>
<th>Time 2: large change, retain position</th>
<th>Time 2: small change, change position</th>
<th>Time 2: small change, change position</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>90</td>
<td>91</td>
<td>50</td>
<td>91</td>
<td>92</td>
</tr>
<tr>
<td>91</td>
<td>91</td>
<td>92</td>
<td>51</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>92</td>
<td>92</td>
<td>93</td>
<td>52</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>93</td>
<td>93</td>
<td>94</td>
<td>53</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>$r = 1.0$</td>
<td>$r = 1.0$</td>
<td>$r = 1.0$</td>
<td>$r = 1.0$</td>
<td>$r = .6$</td>
<td>$r = 0.0$</td>
</tr>
</tbody>
</table>
repeated assessments to start including means and standard deviations of scores at all time points, means and standard deviations of change scores, and correlations between scores at all time points.) But when confronted with multiple options, which reliability coefficients should you choose? For example, if you are repeating the California Verbal Learning Test-II, stability coefficients for Long-Delay Free Recall are presented in the test’s manual (Delis, Kramer, Kaplan, & Ober, 2000; \( r = .88 \)), as well as in published literature (Benedict, 2005; \( r = .54 \); Woods, Delis, Scott, Kramer, & Holdnack, 2006; \( r = .83 \)). As with choosing normative data, a general rule of thumb for choosing reliability values would be to choose the study that best matches your individual patient. This may mean that a clinician utilizes different reliability values when evaluating change in older versus younger patients, less-educated versus more-educated patients, and traumatic brain injury versus Multiple Sclerosis patients.

**Practice effects.** On repeat testing, improvements can occur due to natural recovery or intervention, but improvements can also occur due to prior exposure to the testing materials, and these latter improvements are typically referred to as practice effects. The improvements due to practice effects are probably related to both declarative (e.g., remembering the actual items on the tests) and procedural (e.g., remembering how to do the test) memory and perhaps other cognitive domains (e.g., intelligence, executive functioning). Practice effects are one of the most widely investigated phenomena in serial assessments in neuropsychology, as researchers and clinicians try to identify how much change is normally expected on retesting. Much of this research has shown that practice effects are not uniform across neuropsychological measures; some tests show minimal learning effects, whereas others show large learning effects. For example, on repeat administration of the WAIS-IV, participants improve very little on the Vocabulary and Comprehension subtests (+0.1 and +0.2 scaled score points, respectively, Table 4.5 of Technical and Interpretive Manual). Conversely, more sizable improvements were observed on retesting with the Picture Completion and Visual Puzzles subtests (+1.9 and +0.9 scaled score points, respectively). Presumably, the smaller practice effects occur on subtests that are less novel, ones based on crystallized abilities, where answers are either known or not, and where the responses are previously well-rehearsed (e.g., in school settings). The larger practice effects seem to occur on subtests that are more novel, ones based on fluid abilities, where answers can be acquired in the setting, and where the responses have not been encountered previously. Although clinical lore tends to be contrary, much of the empirical literature tends to support that practice effects:

(a) can occur even if the retest interval is longer than 6 months;
(b) remain relevant even with high test–retest reliability;
(c) are present in children;
(d) are present in older adults; and
(e) are present in patients with a variety of neuropsychological conditions.

Additionally, despite considerable effort in trying to minimize the systematic error associated with these artificial improvements on retesting, some recent research suggests that practice effects may have clinical utility. In three separate clinical samples (Mild Cognitive Impairment [MCI], Human Immunodeficiency Virus, Huntington’s disease), practice effects predicted longer-term cognitive outcomes, above and beyond the baseline test scores (Duff et al., 2007). In other samples of MCI, practice effects have provided useful diagnostic information (Darby, Maruff, Collie, & McStephen, 2002; Duff et al., 2008). Lastly, practice effects have predicted treatment response to a memory training course in older adults (Calero & Navarro, 2007; Duff, Beglinger, Moser, Schultz, & Paulsen, 2010). So, despite largely being viewed as error that needs to be controlled, practice effects may have some diagnostic, prognostic, and treatment implications.

**Novelty.** Related to practice effects are novelty effects. During an initial evaluation, most neuropsychological tests are novel to the patient. However, on repeat testing, these measures may become more familiar. But does that familiarity improve performance or worsen it? Although understudied, the effects of novelty seem equivocal. Whereas some have found that novel tasks improve performance (Kormi-Nouri, Nilsson, & Ohta, 2005), others have found that familiar tasks enhance performance (Poppenk, Kohler, & Moscovitch). It is possible that novelty on initial testing leads to decrements in performance, but familiarity (or release from novelty) on retesting leads to improved performance. In a twist on this theme, Suchy, Kraybill, and Franchow (2011) have found that individuals who do not respond well in novel situations are at greater risk for the cognitive decline. So even though there might still be much to learn about novelty effects, the limited literature suggests that it could be both a confounding variable in repeat assessments and a marker of disease progression, similar to practice effects.

**Floor and ceiling effects.** Floor effects refer scores at or close to the lowest level of performance. Ceiling effects refer to the opposite extreme (i.e., scores at or close to the highest level of performance). In repeat assessment cases, both of these extremes could factor into the amount of change that is possible. For example, if a patient’s performance on the Delayed Recall trial of
the Hopkins Verbal Learning Test-Revised is zero (raw score) at baseline, then the opportunity to find decline is hampered by floor effects. Conversely, if you are looking for benefits of cognitive rehabilitation in a patient with a score of 59/60 correct on the Boston Naming Test, then you are unlikely to find much due to ceiling effects. Therefore, it is important to consider a baseline test score when trying to find change in that score on follow-up. However, it should be noted that floor and ceiling effects are related to scores or scales on tests, and not necessarily to performance or abilities. That is, just because test scores cannot decline further because of floor effects do not mean that this patient cannot worsen across time in his/her abilities.

Variables Associated with the Testing Situation

Retest interval. As noted earlier, the retest interval can affect the reliability of scores across that period. In general, shorter retest intervals lead to higher reliability coefficients, and longer retest intervals lead to lower reliability coefficients. As also alluded to earlier, longer retest intervals can diminish, but not necessarily eliminate, practice effects. So, the amount of time that passes between a baseline and a follow-up appointment is a relevant variable in repeated neuropsychological evaluations. What is the optimal retest interval? As aptly noted in a position paper on serial neuropsychological assessment (Heilbronner et al., 2010), there is insufficient empirical data to develop guidelines on the minimal (or maximal) retest interval in clinical or forensic cases. Even though the decisions about when to retest might be made based on clinical necessity, institutional restrictions, or convenience, the clinician must use his/her knowledge to interpret changes across those intervals.

Regression to the mean. On re-evaluation, a given test score for an individual patient will drift toward the population mean for that test score. For example, a patient with a low score at Time 1 (e.g., Wechsler Memory Scale-IV Logical Memory I demographically corrected T-score = 40) will tend to improve at Time 2 (e.g., T-score = 44) to get closer to the population mean (i.e., T-score = 50). Although some of this improvement could be due to practice and novelty effects, from a statistical standpoint, some is also expected to be due to regression to the mean. In cognitively stable patients, regression to the mean is more evident when high scores at Time 1 drift down (again toward the population mean). For example, a Time 1 T-score of 65 could drop to a T-score of 61 at Time 2 due to these effects. In general, the more extreme score is at baseline, the more likely that regression to the mean effects will occur. However, clinicians need to also be aware of changes that defy these regression effects. For example, the deviant score at baseline that remains stable or gets more deviant at follow-up (e.g., T-score of 40 that drops to 35, T-score of 60 that climbs to 65) probably indicates more change than is actually reflected in raw observed scores, as the score becomes more deviant despite regression to the mean effects.

Variables Associated with the Individual Patient

Demographic variables. Since age, education, gender, and other demographic variables can affect test scores at a single-point evaluation, it is expected that they will exert at least as much of an effect across two assessments. For example, Table 2 shows the amount of change on retesting on the WAIS-IV Block Design subtest across four age groups. Clearly, younger subjects improve more across time than older adults. In another example, Rapport, Brines, Axelrod, and Theisen (1997) found that those with low IQ scores showed smaller practice effects on repeat IQ testing than those with average and high IQ scores. These authors also found that the “rich get richer” on memory tests (Rapport et al., 1997). Although IQ might not be normally viewed as a demographic variable, it does seem related to education, cognitive reserve, and other individual difference variables that affect retesting.

Clinical condition. To follow the reasoning relating to demographic variables, since clinical conditions can affect test scores on a single neuropsychological evaluation, it might be expected that this effect would be compounded with repeated testing. In certain clinical scenarios, we might expect to see effects of the same condition present at both evaluations, albeit at a more severe stage (e.g., Alzheimer’s disease, Huntington’s disease, progressive Multiple Sclerosis). However, in other scenarios, we might see the effects of two different conditions being present at the different evaluations (e.g., psychiatric illness

Table 2. Change on scaled scores on WAIS-IV Block Design across age groups from Wechsler (2008)

<table>
<thead>
<tr>
<th>Age groups</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 2 − Time 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>16–29-year olds</td>
<td>10.1 (3.0)</td>
<td>11.3 (2.9)</td>
<td>+1.2</td>
</tr>
<tr>
<td>30–54-year olds</td>
<td>10.4 (2.9)</td>
<td>11.4 (3.1)</td>
<td>+1.0</td>
</tr>
<tr>
<td>55–69-year olds</td>
<td>10.5 (3.1)</td>
<td>11.1 (2.9)</td>
<td>+0.6</td>
</tr>
<tr>
<td>70–90-year olds</td>
<td>10.0 (2.6)</td>
<td>10.4 (2.5)</td>
<td>+0.4</td>
</tr>
</tbody>
</table>
Practice effects have been observed on cognitive testing as far out as 2.5 years (Salthouse, 2010). Therefore, lengthening a retest interval does not appear to adequately control for repeat testing effects. Prior experiences. Neuropsychologists realize that their patients come to the evaluation with pre-existing strengths and weaknesses based on prior experiences. These strengths can affect test performances on both the initial and follow-up evaluations. For example, Dirks (1982) showed that relatively brief experiences with a commercially-available game would lead to significant improvements on the Block Design subtest of the Wechsler Intelligence Scale for Children-Revised. In this age of video and computer games, patients’ pastimes might be altering their performance, as they introduce “interventions” before or between assessments. Although one cannot control for all possible prior experiences that might influence testing, a thorough clinical interview can identify some of the more likely ones.

Methods for Assessing Change

When working with an individual patient and planning a re-evaluation, a clinician has a host of methodological practices to consider that may allow him/her to make more accurate interpretations of change. These methodologies can be applied to the testing situation to try and minimize the effects of repeated assessments. Additionally, statistical techniques can be used to determine if the observed changes are reliable and clinically meaningful.

Methods Associated with the Testing Situation

Retest interval. As noted earlier, alterations in the retest interval can affect reliability and practice effects on a follow-up visit. However, as also noted earlier, there is limited evidence to identify an optimal retest interval in clinical and forensic cases. Practice effects have been observed on cognitive testing as far out as 2.5 years (Salthouse, 2010). Therefore, lengthening a retest interval does not appear to adequately control for repeat testing effects.

Alternate forms. Several widely used neuropsychological measures have alternate forms that might be appropriate for serial testing. For example, both the Hopkins Verbal Learning Test-Revised and the Brief Visuospatial Memory Test-Revised have six alternate forms available. But it is also obvious that many other widely used measures do not have well-validated alternate forms, including those in the Wechsler intelligence and memory scales, Halstead–Reitan Battery, and most aphasia batteries. Additionally, even existing alternate forms might not be ideal (e.g., identical test format, comparable but different test content, identical psychometric properties). For example, despite have six alternate forms, all of the alternate forms of the Hopkins Verbal Learning Test-Revised do not appear to be comparable (Benedict, Schretlen, Groninger, & Brandt, 1998). Furthermore, alternate forms do not guarantee that practice effects will not occur. Beglinger and colleagues (2005) have demonstrated practice effects on serial testing when alternate forms were used.

Appropriate control groups. In research studies, the inclusion of a control group, especially in longitudinal studies, significantly improves the scientific value of the study. “Normal” cognitive change in a control group (i.e., not affected by the intervention of interest) can be compared with the cognitive change in an experimental group to better evaluate the effects of the intervention. In most research studies, subjects are randomly assigned to either the experimental or a control group, which increases the chances that these two groups will be comparable (except for the intervention). However, when working with an individual patient, a clinician does not have the opportunity to assign a similar patient to a control group to look for “normal” change. This clinician must look to the existing literature to find studies that match his/her patient in demographics, retest interval, and neuropsychological measures. The more that a study’s sample matches the individual patient, the more that this study can be used for “change norms” for this individual patient. An initial question that might arise is: how much must the sample characteristics match the individual patient? For example, must they be identical for age, education, gender, and retest interval? Just as clinicians can struggle to find normative data (for a single assessment) that exactly matches their individual patients, finding change norms can be even more of a challenge. Each clinician will have to decide how close is close enough, and then account for any notable discrepancies in the interpretation of the data. A second likely question might be: is it better to find change norms on healthy controls or those with a similar diagnosis? Surprisingly, the literature contains many more examples of “clinical change norms” and less examples of change in cognitively healthy samples. But it is likely that these two sets of norms, if they can be located, will complement one another. Change norms in healthy individuals will indicate if the amount of change observed in the individual patient differs significantly from that seen in healthy persons (e.g., is this amount of change more than expected in “normal” individuals?). Change norms in diagnostically similar samples will indicate if the amount of change observed in the individuals differs from that diagnostic group (e.g., is this amount of change more than
expected in other patients with medulloblastomas?). Implied earlier is a third likely question: can I access these change norms? Unfortunately, there are no standards or guidelines for reporting serial assessment data in empirical articles or test manuals, and many such reports exclude some of the key elements for determining change across time. At a minimum, it is necessary to have baseline and follow-up means and standard deviations for test scores, as well as test–retest reliability coefficients. Means and standard deviations of change scores (e.g., Time 2 – Time 1) are also helpful. With this information, most reliable change indexes (RCIs; below) can be calculated.

Methods for Assessing Reliable Change

There are several statistical methods that are used to assist the clinician in determining if a reliable change has occurred across time. The formulas for these different methods are presented in Table 3. In the examples below, \( T_1 \) = score at Time 1, \( T_2 \) = score at Time 2, \( M_1 \) = mean score of control group at Time 1, \( S_1 \) = standard deviation of control group at Time 1, \( M_2 \) = mean score of control group at Time 2, \( S_2 \) = standard deviation of control group at Time 2, \( r_{12} \) = correlation between \( M_1 \) and \( M_2 \). Additionally, for most of the examples below, we will use the following hypothetical scores (standard scores with \( M = 100 \) and \( SD = 15 \)) and psychometric properties: 

- \( T_1 = 90 \)
- \( T_2 = 80 \)
- \( M_1 = 100 \)
- \( S_1 = 15 \)
- \( M_2 = 105 \)
- \( S_2 = 20 \)

and \( r_{12} = .85 \).

Simple discrepancy score. Perhaps the most intuitive of all methods for evaluating change between two testing scores is the simple discrepancy score. This discrepancy score is calculated as the difference between Time 1 and Time 2 scores (Table 3). This discrepancy score is then compared with normative data, which will show the frequency of this discrepancy score in some sample. On the positive side, the simple discrepancy score might be the easiest one to calculate. On the negative side, the clinician needs access to the normative data of discrepancy scores in a relevant sample. Additionally, this simple discrepancy method is expected to be a less precise estimate of relative change because the clinician is often left with a range of values. It is also a one-size-fits-all approach and does not specifically control for factors known to affect repeated assessments (e.g., varying ages, retest intervals).

Patton and colleagues (2005) provides an example of the simple discrepancy score. In this study, the authors generated base rates of discrepancy scores for a healthy elderly sample using the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Table 4). In our patient example, the simple discrepancy would be \(-10\) (i.e., \(80–90\)). Using Table 4 (which

<table>
<thead>
<tr>
<th>Table 3. Reliable change scores and their formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reliable change score</strong></td>
</tr>
<tr>
<td>Simple discrepancy score</td>
</tr>
<tr>
<td>Standard Deviation Index</td>
</tr>
<tr>
<td>RCI</td>
</tr>
<tr>
<td>RCI controlling for practice effects (RCI + PE) and alternate calculation of SED by Iverson</td>
</tr>
<tr>
<td>SRB</td>
</tr>
<tr>
<td>Estimated SRB calculated from retest data</td>
</tr>
</tbody>
</table>

Notes: \( T_1 = \) score at Time 1; \( T_2 = \) score at Time 2; \( S_1 = \) standard deviation at Time 1; \( S_2 = \) standard deviation at Time 2; \( r_{12} = \) correlation between Time 1 and 2 scores; \( M_1 = \) control group mean at Time 1; \( M_2 = \) control group mean at Time 2; \( b = \) slope of the regression model (beta coefficient); \( c = \) intercept of the regression model (constant); \( SED_{\text{est}} = \) standard error of the estimate of the regression model; \( T_2' = \) predicted score at Time 2 based on the regression model; \( RCI = \) Reliable Change Index; \( PE = \) practice effect; \( SED = \) standard error of the difference; \( SRB = \) standardized regression-based formula.

<table>
<thead>
<tr>
<th>Table 4. Simple discrepancy score from Patton and colleagues (2005)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cumulative percentages</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Age-Corrected Total Scale (OKLAHOMA1-year interval [T2 – T1])</td>
</tr>
</tbody>
</table>
coincidentally is also Table 4 from Patton et al.) and assuming this is an age-corrected Total score from the RBANS (OKLAHOMA norms, 1-year retest interval), this discrepancy falls between −11 (10%) and −8 (20%) of that sample. Therefore, you could conclude that the amount of change observed in the example patient occurs in 10%–20% of a healthy elderly sample.

**Standard Deviation Index.** Whereas the simple discrepancy method might be the easiest change method to use, the Standard Deviation Index might be one of the most widely used among clinicians. In this method, the simple discrepancy score is divided by the standard deviation of the test score at Time 1. This yields a $z$-score, which can be compared with a normal distribution table to find out the statistical significance of that difference. Within the existing literature, a $z$-score of ±1.645 would typically be considered a “reliable change.” This ±1.645 demarcation point indicates that 90% of change scores will fall within this range in a normal distribution and only 5% of cases will fall below this point based on chance and only 5% of cases will fall above this point. One advantage of the Standard Deviation Index is that it is easy to calculate. It also provides a more precise estimate of relative change than the simple discrepancy score because it is tied to a specific $z$-score. Disadvantages associated with this method include: no control for test reliability, practice effects, or regression to the mean, and it is a one-size-fits-all approach. Additionally, as it puts change on a scale of standard deviation units, it is quantifying change on an incorrect metric (as will be described with the following methods).

In our patient example, the Standard Deviation Index would be $-0.67$ (i.e., $(80 - 90)/15)$. When compared with a normal distribution table, a $z$-score of −0.67 falls at approximately the 25th percentile. Since this falls well above the typical cutoff of ±1.645, then a clinician would conclude “no change.” When one compares the simple discrepancy score (roughly 10th to 20th percentile) and the Standard Deviation Index (25th percentile), it is apparent that they are close, but not identical. Since the simple discrepancy score is tied to actual changes in some normative group, it is likely to be more accurate reflection of change in the individual patient than the standard deviation index, which is tied to psychometric properties of the test from a single administration (e.g., standard deviation at Time 1). However, in the absence of access to any better methods, the Standard Deviation Index is favorable to a clinician’s best guess about change.

**Reliable Change Index.** First developed to determine if clinically meaningful change occurred as a result of psychotherapy (Jacobson & Truax, 1991), the RCI is a more sophisticated method for examining change. Similar to the standard deviation index, it uses the simple discrepancy between the Time 1 and Time 2 scores as the numerator. But unlike the standard deviation index, it uses the standard error of the difference (SED) in the denominator. In essence, the SED estimates the standard deviation of the differences scores (which is likely to be very different than the SD of Time 1 scores used in the SD index). Although the SED continues to include the standard deviation at Time 1, it also incorporates the reliability of the test (Table 3). This makes the RCI a notable advancement over the prior two methods. Calculation of the RCI results in a $z$-score similar to the standard deviation index, which needs to be compared with a normal distribution table. Advantages of the RCI include: a more precise estimate of relative change and control for the test’s reliability. Disadvantages include: it does not correct for practice effects or variability in Time 2 scores and it remains a one-size-fits-all approach.

In the patient example, the RCI’s numerator would also be −10 (i.e., 80−90). The RCI’s denominator would be 8.22 (i.e., SED = $\sqrt{2 \times 15^2(1 - 0.85)}$). This would result in an RCI of −1.22 (i.e., −10/8.22). Compared with a normal distribution table, a $z$-score of −1.22 falls at approximately the 12th percentile. Since this falls above our typical cutoff of ±1.645, then you would conclude “no change.” Despite finding “no change,” the accuracy of the RCI is noticeable compared with the other two methods, which is attributable to the additional error variance that is controlled for in the denominator of this method.

**RCI + practice effects.** Although the RCI was a notable improvement in assessing change, it was designed for measures of psychological constructs (e.g., depression, anxiety). Cognitive measures, however, change differently than psychological measures. In particular, many cognitive measures show practice effects on repeat testing, which is not accounted for in the RCI method. Therefore, Chelune, Naugle, Luders, Sedlak, and Awad (1993) adjusted the RCI to control for practice effects (RCI_{PE}). The numerator of RCI_{PE} starts with the simple discrepancy score (i.e., Time 2 − Time 1). From this, discrepancy score is subtracted the mean practice effects from some relevant group (which could be healthy controls or a clinical sample). This practice-adjusted discrepancy score is the numerator in RCI_{PE}. In their original paper, Chelune and colleagues used the SED as the denominator. The resulting RCI_{PE} is compared with a normal distribution table, and ±1.645 is also used as a cutoff point for considering a statistically significant change. In addition to being a more precise estimate of relative change and controlling for the test’s reliability, the main advantage of RCI_{PE} is that it controls for practice effects. One disadvantage of the RCI_{PE} method is that the practice effects correction is uniform (i.e., it does not allow for differential practice effects). Additionally, it remains a one-size-fits-all approach and does not control for variability in Time 2 scores.
In our patient example, the numerator of our RCIPE would be −15 (i.e., (80−90)−(105−100)). The denominator would still be 8.22 (i.e., SED = \sqrt{2 \times 15^2(1−0.85)}). The resulting RCIPE would be −1.83 (i.e., −15/8.22). Compared with a normal distribution table, a z-score of −1.83 falls at approximately the 4th percentile. Since this value falls below our typical cutoff of ±1.645, then you could conclude that there had been a reliable and meaningful “change.”

Although the SED had been used for some time, Iverson (2001) observed that the variability in the Time 2 scores was not unaccounted for in existing formulas. He introduced an adapted SED that does incorporate Time 2’s variability (SED_{Iverson}), and this alternate calculation is now typically used as the denominator in RCIPE. In our patient example, the numerator remains −15. The denominator changes to 9.68 (i.e., SED_{Iverson} = \sqrt{(15^2(1−0.85)) + (20^2(1−0.85))} = \sqrt{(5.81^2 + (7.4)^2)} = \sqrt{93.67}), and the RCIPE is now −1.55 (approximately 6th percentile but “no change” according to ±1.645).

A few observations are probably necessary at this point. First, even though the previous methods might differ in the exact point at which this change score is located (e.g., 10th – 20th for simple discrepancy, 25th for standard deviation index, 12th for RCI, 4th for RCIPE, 6th for RCIPE with SEDIverson), they all consistently indicate some trend toward a decline in scores (i.e., all fall on the lower end of the distribution). Second, as more information is added to the equation, including test reliability, practice effects, and variability at Time 1 and Time 2, the estimate of change improves in accuracy. Third, the point at which we decide “change/no change” (i.e., ±1.645) is somewhat arbitrary, as many other factors must be considered when interpreting neuropsychological test scores. Lastly, all of the previous methods are constrained because they are unidimensional and rigid. This one-size-fits-all approach to assessing change does not account for differences in the individual patient (e.g., age, education, baseline level of performance, differential practice effects).

**Regression-based change formulas.** Developed around the same time (and by some of the same authors) as the RCIPE was a regression-based method for determining if meaningful cognitive change had occurred (McSweeny, Naugle, Chelune, & Luders, 1993). This method utilized multiple regression to predict a Time 2 score using the Time 1 score and other possibly relevant clinical information (e.g., age, education, retest interval). In the original McSweeny and colleagues paper, only the Time 1 score was a significant predictor of the Time 2 score (i.e., no other variables entered the equation), and we refer to these as “simple” standardized regression-based formulas (simple SRB). With this method, a predicted Time 2 score could be generated in $T_2 = bT_1 + c$, where $T_2$ is the predicted Time 2 score, $b$ the $\beta$ weight for Time 1 score (or regression slope), $T_1$ the Time 1 score, and $c$ the constant (or regression intercept). The predicted score could then be tested in $RCI_{SRB} = T_2 − T_2/SEE$, where $SEE$ is the standard error of the estimate of the regression equation. The resulting $RCI_{SRB}$ also needs to be compared with a normal distribution table, and ±1.645 is again used as a typical cutoff point for considering change. Unlike its predecessors, the SRB model does allow for other variables in the prediction of a Time 2 score. In the case of the simple SRB, Time 1 cognition is accounted for in the model. This may be important if the Time 1 score falls at one extreme or another (e.g., high Time 1 scores may show less improvement on retesting due to ceiling effects, low Time 1 scores may show less decline on retesting due to floor effects). Additionally, regression to mean affects scores differently depending on their starting point (e.g., high Time 1 scores are more likely to regress downward, low Time 1 scores are more likely to regress upward). Other advantages of the simple SRB are that it provides a more precise estimate of relative change, it corrects for practice effects and retest reliability, and it corrects for variability in Time 2 scores. Furthermore, the SRB method can potentially incorporate additional clinically relevant variables (e.g., age, education, retest interval) into the prediction model, and we refer to this as the “complex” SRB approach. Although McSweeny and colleagues did not find that other variables to significantly contributed to the prediction of Time 2, more recent studies have found that demographic variables and retest interval contribute small, but statistically significant, amounts of variance for certain cognitive measures. Disadvantages of the SRB approach have primarily centered on that these formulas are complicated to calculate. Additionally, unless these formulas are already published, one would need access to an appropriate sample with test–retest data to generate the necessary regression analyses.

To continue with our patient example, we utilized the published simple SRB for the Repeatable Battery for the Assessment of Neuropsychological Status in older adults retested after 1 year (Duff et al., 2004). Using Table 5, the Time 2 Delayed Memory Index is best predicted by the Time 1 score on that same measure (i.e., 90) multiplied by the $\beta$ coefficient (i.e., 0.71) plus the constant (i.e., 30.60), yielding a $T_2' = 94.5$ (i.e., $T_2' = 90 \times 0.71 + 30.60$). The $T_2'$ is subtracted from the $T_2$ and divided by the SEE of the regression equation, to yield an $RCI_{SRB} = -1.26$ (i.e., $T_2 − T_2'/SEE = 80 − 94.5/11.46$). Compared with a normal distribution table, a z-score of −1.26 falls at approximately the 10th percentile. Since this falls above our typical cutoff of ±1.645, then you would conclude “no change.” If other variables were included in the regression models, such as the Immediate Memory Index in Table 5, then this is a complex SRB (e.g., age and education add to the prediction of the Time 2 score).

One criticism of the SRB approach is that you typically need access to the actual data of relevant samples to generate the regression analyses. However, two groups have demonstrated that the key elements of the $RCI_{SRB}$ can be estimated from...
psychometric properties that are typically available in test manuals and published reports (Crawford & Garthwaite, 2007; Maassen, Bossema, & Brand, 2009). For example, with means and standard deviations at Time 1 and Time 2 from a relevant sample and the test–retest reliability coefficient, one can calculate a simple SRB and related $\text{RCI}_{\text{SRB}}$ (Table 3). Whereas the constant and $\beta$ coefficient used to calculate $T_2'$ would normally be taken from the regression results, they can be estimated from the means and standard deviations at Time 1 and Time 2 for a relevant sample. Similarly, the $\text{SEE}$, which would normally be taken from the regression analyses, can be estimated from the standard deviations at Time 1 and Time 2 and the test’s reliability. The final calculation of this estimated $\text{RCI}_{\text{SRB}}$, which we label $\text{RCI}_{\text{SRBest}}$, is similar to that coming directly from the regression analyses (i.e., $T_2 - T_2'/\text{SEE}$).

In our patient example, $T_2'$ would be 91.67 (i.e., $b_{\text{est}} = 20/15 = 1.33$; $c_{\text{est}} = 105 - 1.33 \times 100 = -28$; $T_2' = 1.33 \times 90 - 28 = 91.7$). The $\text{SEE}_{\text{est}}$ would be 9.68 (i.e., $\text{SEE}_{\text{est}} = \sqrt{(S_{T_1}^2 + S_{T_2}^2)(1 - r_{12})} = \sqrt{15^2 + 20^2}(1 - 0.85) = 9.68$). The $\text{RCI}_{\text{SRBest}}$ would be $-1.21$ (i.e., $\text{RCI}_{\text{SRBest}} = T_2 - T_2'/\text{SEE}_{\text{est}} = 80 - 91.7/9.68 = 1.21$). Compared with a normal distribution table, a $z$-score of $-1.21$ falls at approximately the 12th percentile. Since this falls above our typical cutoff of $\pm 1.645$, then you would conclude “no change.”

There are additional variations on these different statistical methods for examining change. For example, Crawford and Garthwaite (2006) noted that an adjustment is needed to the denominator in SRBs to control for a new case. Additionally, RCIs have been calculated for entire batteries, not just individual measures (Woods, Childers, et al., 2006). Various debates have tried to refine these methods and identify instances when one is preferred to another (Hinton-Bayre, 2005, 2010; Maassen, Bossema, & Brand, 2006). This final debate is one worth briefly addressing: which change formula is best?

A number of authors have compared various RCI methods to determine their effectiveness in identifying change. Temkin, Heaton, Grant, and Dikmen (1999) compared four of these methods (RCI, $\text{RCI}_{\text{PE}}$, simple SRB, and complex SRB) in a large sample of neurologically stable adults on five measures and two summary scores from the Halstead–Reitan Neuropsychological Test Battery. Results indicated that the original RCI was the poorest at identifying change, but that the other three methods were largely comparable. Two years later, Heaton and colleagues (2001) examined the $\text{RCI}_{\text{PE}}$, simple SRB, and complex SRB in non-clinical and clinical samples on the same cognitive variables examined by Temkin and colleagues. Again, all three methods were found to be comparable, and it was noted that change models in normals might not apply to clinical cases. Frerichs and Tuokko (2005) compared the standard deviation index, RCI, $\text{RCI}_{\text{PE}}$, simple SRB, and complex SRB in a large cohort of cognitively normal seniors on four memory measures. Results found greatest agreement between the $\text{RCI}_{\text{PE}}$, simple SRB, and complex SRB.

Table 5 provides references for many of these relevant studies.
The assessment of cognitive change in the individual patient will remain an important component of a neuropsychologist’s job responsibilities in the future. Although this part of clinical neuropsychology has grown rapidly over the past 20 years, there is still much room for additional growth. Some important future directions include the following.

- Examining these methods in geriatric and pediatric samples. Although there is a wealth of existing data on reliable change in adult samples (both controls and clinical cases), there is a dearth of relevant information on those under 18 and over 65 years of age. These two opposite ends of the age spectrum have unique developmental and degenerative processes that may make adulthood change norms less applicable.

- Better coverage of methods in clinical samples. Although some clinical conditions have been better studied with RCIs and SRBs (e.g., epilepsy, Parkinson’s disease), others are woefully under-represented (e.g., Multiple Sclerosis, dementia, traumatic brain injury, brain tumors). Presumably, these under-represented conditions are being seen for repeated neuropsychological evaluations, but clinicians are not compiling this data, calculating these change indexes, and/or publishing their findings. We implore them to do so.

- Who is the ideal comparison group? When evaluating a patient with a traumatic brain injury for a repeat evaluation, is it best to compare his/her change to cognitively healthy controls? Or should his/her performance be compared with others with similar traumatic brain injuries? As noted earlier, both types of comparisons likely yield valuable information. However, Heaton and colleagues (2001) opined that “normal” change might not be applicable in clinical cases. To our knowledge, no one has empirically evaluated this assumption. If Heaton is correct, then it is even more critical that we increase our research efforts on determining what amount of change is expected in various disease states.

- Should raw scores be used to determine reliable change? Or corrected scores? In their original paper on SRBs, McSweeny and colleagues (1993) actually used a mix of raw and corrected scores in their analyses of change on the Wechsler Memory Scale-Revised and the WAIS-Revised in patients with epilepsy. Their argument for using raw scores with the Wechsler Memory Scale-Revised was that it led to a better fit of the data, and their argument for using corrected scores with the WAIS-Revised was that the age-corrected IQ scores would be more understandable to their audience. Regardless of one’s arguments/choices, a consumer of RCIs and SRBs should always use the same metric that was used in the relevant publication. For example, if I want to use McSweeny’s SRBs

**Table 6. Selected citations for studies using RCIs and SRBs in clinical samples**

<table>
<thead>
<tr>
<th>Epilepsy</th>
<th>Parkinson’s disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCI (Chelune et al., 1993)</td>
<td>RCI (Troster, Woods, &amp; Morgan, 2007)</td>
</tr>
<tr>
<td>SRB (McSweeny et al., 1993)</td>
<td>RCI (Higginson et al., 2009)</td>
</tr>
<tr>
<td>Both (Hermann et al., 1996)</td>
<td>RCI (Zahodne et al., 2009)</td>
</tr>
<tr>
<td>Both (Sawrie, Chelune, Naugle, &amp; Laders, 1996)</td>
<td>SRB (Muslimovic, Post, Speelman, De Haan, &amp; Schmand, 2009)</td>
</tr>
<tr>
<td>Both (Martin et al., 2002)</td>
<td>RCI (Mikos, Zahodne, Okun, Foote, &amp; Bowers, 2010)</td>
</tr>
<tr>
<td>Both (Martin, Griffith, Sawrie, Knowlton, &amp; Faught, 2006)</td>
<td>SRB (Rinehardt et al., 2011)</td>
</tr>
<tr>
<td>Both (Nakhutina, Pramataris, Morrison, Devinsky, &amp; Barr, 2010)</td>
<td>RCI (Schoenberg et al., 2011)</td>
</tr>
<tr>
<td>RCI (Loring, Williamson, Meador, Wiegand, &amp; Hullhan, 2011)</td>
<td>Dementia</td>
</tr>
<tr>
<td>Traumatic brain injury</td>
<td>RCI (Hensel et al., 2009)</td>
</tr>
<tr>
<td>Both (Barr &amp; McCrea, 2001)</td>
<td>RCI (Rossetti, Munro Cullum, Hynan, &amp; Lacritz, 2010)</td>
</tr>
<tr>
<td>RCI (Millis et al., 2001)</td>
<td></td>
</tr>
<tr>
<td>RCI (Barr, 2003)</td>
<td></td>
</tr>
<tr>
<td>RCI (Till, Colella, Verwegen, &amp; Green, 2008)</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td></td>
</tr>
<tr>
<td>Both (Portaccio et al., 2010)</td>
<td></td>
</tr>
<tr>
<td>HIV</td>
<td></td>
</tr>
<tr>
<td>Both (Levine et al., 2007)</td>
<td></td>
</tr>
<tr>
<td>SRB (Cysique et al., 2011)</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td></td>
</tr>
<tr>
<td>Both (Portaccio et al., 2010)</td>
<td></td>
</tr>
<tr>
<td>HIV</td>
<td></td>
</tr>
<tr>
<td>Both (Levine et al., 2007)</td>
<td></td>
</tr>
<tr>
<td>SRB (Cysique et al., 2011)</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td></td>
</tr>
<tr>
<td>Both (Portaccio et al., 2010)</td>
<td></td>
</tr>
<tr>
<td>HIV</td>
<td></td>
</tr>
<tr>
<td>Both (Levine et al., 2007)</td>
<td></td>
</tr>
<tr>
<td>SRB (Cysique et al., 2011)</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** RCI = reliable change index; SRB = standardized regression-based formula; both = RCI and SRB data presented.

The assessment of cognitive change in the individual patient will remain an important component of a neuropsychologist’s job responsibilities in the future. Although this part of clinical neuropsychology has grown rapidly over the past 20 years, there is still much room for additional growth. Some important future directions include the following.
for the Wechsler Memory Scale-Revised, then I need to be using raw scores too. However, there is no literature to
guide us on which is actually best when developing these change models.

- Expanding the methodology beyond specific cognitive tests. The vast majority of RCIs and SRBs are developed
for individual neuropsychological test scores. However, future RCI and SRB studies might employ a battery-
wise approach, like done by Woods, Childers, et al. (2006). Additionally, and perhaps more widely applicable,
would be a shift to domain-specific RCIs and SRBs. Duff, Beglinger, Moser, & Paulsen (2010) examined if
SRBs could be generated that predicted Time 2 scores on one test from Time 1 scores on a different test
from the same cognitive domain (e.g., predicting Time 2 scores on Delayed Recall of Hopkins Verbal
Learning Test-Revised from the Time 1 score on List Recall of the Repeatable Battery for the Assessment of
Neuropsychological Status). Although the results were promising (e.g., domain-specific SRBs were comparable
with test-specific SRBs), these results need to be validated and expanded. Furthermore, RCIs and SRBs could
be generated for psychiatric and functional scales, MRI volumes, or other relevant outcome measures when evalu-
ating changes in neuropsychological status.

- Handling more than two testing sessions. Nearly, all studies of cognitive change have examined two times points,
but we are increasingly seeing patients who are being evaluated a third or fourth time. Can you use the same RCIs
and SRBs to compare changes between Times 2 and 3 that you used to compare Times 1 and 2? Probably not, but
there are only a few studies that have provided initial evidence of how cognitive changes vary with multiple
assessments (Attix et al., 2009; Duff, Schoenberg, et al. (2008)). Other statistical methods (e.g., latent growth
curve modeling) may be more appropriate for these complex trajectories.

- Refining methods. Although neuropsychologists have multiple methods at their disposal to assess change, the
variables that go into these equations have not been successful in capturing all of the variance associated
with true change. For example, Martin and colleagues (2002) developed SRBs for the WAIS-III and the
Wechsler Memory Scale-III in a sample of non-operated epilepsy patients, and the resulting equations captured
31%–92% of the variance, even though baseline test score, age, gender, and seizure information were included
as predictor variables. And these results reflect better-than-average SRBs. Therefore, we need to identify add-
itional variables that might increase the captured variance in change models, perhaps including quality of edu-
cation, premorbid intellect, medical and psychiatric information, occupational status, and performance in other
cognitive domains.

- Overcoming obstacles for implementation in clinical practice. One potential reason for underutilization of
change formulas by clinicians (and researchers) is that these formulas are cumbersome to calculate. Following the lead of Dr. Crawford (see http://www.abdn.ac.uk/~psy086/dept/psychom.htm), we have become
advocates for providing interested readers with change score calculators (e.g., Microsoft Excel spreadsheets)
of our relevant work in this area. Interested readers can contact the first author for an example of one such cal-
culator. We also strongly encourage other authors to follow this model.

- How should reliable change be addressed in forensic cases? Besides clinical cases, another venue where repeated
assessment is common is in forensic evaluations. In an extreme case, a personal injury case that was tested by two
different neuropsychologists on two successive days (Putnam, Adams, & Schneider, 1992). Although both eval-
uations produced comparable opinions, notable practice effects were observed across several measures, which
could affect data interpretation. In another example, O’Mahar and colleagues (in press) recently reported that the
1-year test–retest stability of the Effort Index of the Repeatable Battery for the Assessment of
Neuropsychological Status was relatively low (e.g., $r = .32–.36$) in two samples of geriatric patients. The reli-
ability and reliable change observed on other effort measures has been notably understudied. In general, neurop-
sychologists should attempt to inform the courts about the potential complications of repeated evaluations and
interpret their data accordingly (Heilbronner et al., 2010). However, more guidance and empirical data is clearly
needed to assist neuropsychologists in forensic cases with repeated assessments.

- Is $± 1.645$ the best cutoff for determining change? Although this demarcation point was originally chosen
because of its parallel with traditional parametric statistical testing, there is little (if any) data to support it as
the best cut-point for assessing change. Improvements of $± 1.53$ or declines of $± 1.18$ still tell us something
about change, even though they fall within the “no change” range.
What is true change? Despite RCI scores, there are probably real-life events that also indicate change. When a patient with a traumatic brain injury can return to work, then change has probably occurred. When a slowly dementing patient can no longer live alone, change has occurred. When seizures become so disruptive that surgery is sought, change has occurred. When a child with Attention Deficit Hyperactivity Disorder shows improving grades in school while taking a stimulant medication, change has occurred. Although we currently track change with test scores, we probably need to be examining how our test scores track with real-life indicators of change.

In conclusion, repeated assessment is a relatively common occurrence in clinical neuropsychology that carries distinct benefits and unique challenges. Neuropsychologists have a variety of choices to make, both methodologically and statistically, when trying to determine if significant, reliable, and meaningful change has occurred. Despite the growing popularity of serial assessments and the expanding literature in this area, there is a need for more empirical studies to address several important but unanswered questions. We encourage those with relevant data to publish their findings to further inform the field.

Funding

The project described was supported by research grants from the National Institutes on Aging (K23 AG028417) to KD.

Conflict of Interest

None declared.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health. Portions of this article were presented at the 2010 Annual Conference of the National Academy of Neuropsychology, Vancouver, BC.

References


