OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 35 (2020) 1168–1181

Archives of CLINICAL NEUROPSYCHOLOGY

Development and Preliminary Validation of Standardized Regression-Based Change Scores as Measures of Transitional Cognitive Decline

Andrew M. Kiselica[1,†,*], Alyssa N. Kaser[2], Troy A. Webber[3], Brent J. Small[4], Jared F. Benge[1,5,6]

[1]*Division of Neuropsychology, Baylor Scott and White Health, Temple, TX, USA*
[2]*Department of Psychology and Neuroscience, Baylor University, Waco, TX, USA*
[3] *Michael E. Debakey VAMC, Houston, TX, USA*
[4]*School of Aging Studies, University of South Florida, Tampa, FL, USA*
[5] *Plummer Movement Disorders Center, Temple, TX, USA*
[6] *Texas A&M College of Medicine, Bryan, TX, USA*

*Corresponding author at: Department of Neurology, Division of Neuropsychology, Baylor Scott and White Health, 2301 S 31st Street, Temple, TX 76508, USA. Tel.: 215-962-5842. *E-mail address*: andrew.kiselica@bswhealth.org (A.M. Kiselica).

Editorial Decision 11 June 2020; Accepted 11 June 2020

Abstract

Objective: An increasing focus in Alzheimer's disease and aging research is to identify transitional cognitive decline. One means of indexing change over time in serial cognitive evaluations is to calculate standardized regression-based (SRB) change indices. This paper includes the development and preliminary validation of SRB indices for the Uniform Data Set 3.0 Neuropsychological Battery, as well as base rate data to aid in their interpretation.
Method: The sample included 1,341 cognitively intact older adults with serial assessments over 0.5–2 years in the National Alzheimer's Coordinating Center Database. SRB change scores were calculated in half of the sample and then validated in the other half of the sample. Base rates of SRB decline were evaluated at z-score cut-points, corresponding to two-tailed p-values of .20 ($z = -1.282$), .10 ($z = -1.645$), and .05 ($z = -1.96$). We examined convergent associations of SRB indices for each cognitive measure with each other as well as concurrent associations of SRB indices with clinical dementia rating sum of box scores (CDR-SB).
Results: SRB equations were able to significantly predict the selected cognitive variables. The base rate of at least one significant SRB decline across the entire battery ranged from 26.70% to 58.10%. SRB indices for cognitive measures demonstrated theoretically expected significant positive associations with each other. Additionally, CDR-SB impairment was associated with an increasing number of significantly declined test scores.
Conclusions: This paper provides preliminary validation of SRB indices in a large sample, and we present a user-friendly tool for calculating SRB values.

Keywords: Standardized regression-based change; Uniform data set; Norms; Transitional cognitive decline; Assessment

Introduction

The Importance of Assessing Cognitive Change in Serial Evaluations

As interest in capturing and changing the trajectory of cognitive declines with aging and associated neurodegenerative disorders increases (Cummings, 2019; Cummings, Lee, Ritter, Sabbagh, & Zhong, 2019; Jack et al., 2018), the need for reliable

doi:10.1093/arclin/acaa042 Advance Access publication 21 July 2020

methods of interpreting serial cognitive evaluations grows more critical. Indeed, the recent National Institute on Aging and Alzheimer's Association (NIA-AA) research criteria for Alzheimer's disease (Jack et al., 2018) emphasize the importance of capturing subtle cognitive changes preceding the development of diagnoses, such as mild cognitive impairment (MCI) or dementia, via the presence of transitional cognitive decline on objective neuropsychological measures. As the neurodegenerative research world shifts to accommodate this new paradigm, it is crucial for neuropsychologists to create robust tools to quantify these subtle changes over serial evaluations.

It is common in clinical practice to use a simple visual inspection of serial test data to assess for the presence of cognitive deterioration (Lezak, Howieson, & Loring, 2012). However, variability in repeated evaluations can occur due to test, patient, and/or environmental factors, such that interpreting change over time is difficult (Duff, 2012; Hill, 2019). This fact necessitates the development and dissemination of longitudinal normative data and statistical techniques appropriate for distinguishing chance or expected changes from those suggestive of the presence of neurodegenerative disease (Brooks, Sherman, Iverson, Slick, & Strauss, 2011; Brooks, Strauss, Sherman, Iverson, & Slick, 2009b; Heilbronner et al., 2010).

Standardized Regression-Based Approaches

One technique that can effectively identify change in objective cognitive measures is the Standardized Regression-Based (SRB) approach. Complex SRB methods (hereafter simply referred to as SRB) index the probability of a given amount of change in serial evaluations after adjusting for the influence of baseline scores, practice effects, test–retest reliability, and other selected factors of relevance, such as patient demographic characteristics (Calamia, Markon, & Tranel, 2013; Duff, 2012; McSweeny, Naugle, Chelune, & Lüders, 1993). In the SRB approach, the difference between observed follow-up scores and regression-predicted scores is assessed and expressed in standardized format (e.g., as a z-score). In this way, changes over time can be expressed in terms of standard deviation (*SD*) units or converted to percentile ranks. Additionally, different normative cut-points can be assessed to delineate the presence of subtle versus more prominent declines in serial evaluations, which may prove particularly helpful in identifying individuals experiencing subtle but concerning cognitive changes, despite not meeting criteria for MCI or dementia. Regression-based methods have been shown to demonstrate superior clinical utility to other methods, such as reliable change indices (Duff, Suhrie, Dalley, Anderson, & Hoffman, 2019; Hill, 2019; Hinton-Bayre, 2016). Furthermore, they have the advantage of being able to account for relevant demographic variables, unlike reliable change indices (Hill, 2019).

SRB Change Versus Clinically Meaningful Change

The SRB approach is used to assess whether changes over time are statistically likely and expected versus statistically unlikely (Duff, 2012). However, statistical and clinically important changes are not necessarily the same thing. As noted by prior authors, "Significant change does not necessarily equate to clinically significant change. ... Therefore, use of base rate data or effect size to clarify how common/uncommon the degree of change is may be useful." (Hill, 2019, p. 54)

Brooks, Holdnack, and Iverson (2016) demonstrated that statistically unlikely changes are actually quite common when a battery of neuropsychological tests is administered. Approximately one third of healthy older adults in their sample demonstrated at least one statistically unlikely reduction in scores when reassessed at 3–27 weeks. This finding suggests that "abnormal" changes in cognitive performance over time are actually common in normally aging populations. Thus, understanding the base rates of change on measures over time is another critical piece to understand how SRB indices derived from serial cognitive assessments may reflect transitional cognitive decline in older adults.

Change in the Uniform Data Set 3.0 Neuropsychological Battery

As psychometric properties that define change are contingent on the instruments employed and the situations in which these instruments are utilized (Duff, 2012; Hill, 2019), the Uniform Data Set 3.0 Neuropsychological Battery (UDS3NB) provides an excellent means for exploring the questions raised above. The UDS3NB consists of a set of consensus measures administered to patients at Alzheimer's Disease Research Centers across the country (Besser et al., 2018; Weintraub et al., 2018). While developed as a research tool, the UDS3NB has a growing body of evidence to support its use in clinical settings with aging populations (Porto, Russo, & Allegri, 2018) as well as published normative data and interpretive tools (Devora, Beevers, Kiselica, & Benge, 2019; Kiselica, Webber, & Benge, 2020a, 2020b; Kornak et al., 2019; Liew, 2019; Weintraub et al., 2018).

Previous studies calculated reliable change indices and minimally clinically important differences for the UDS 2.0, a past version of the battery (Andrews et al., 2019; Gavett, Ashendorf, & Gurnani, 2015). However, researchers have yet to examine change in the UDS3NB, which represents a largely unique battery of tests (Besser et al., 2018; Kiselica et al., 2020b; Weintraub et al., 2018). Major changes to the battery include: the addition of new measures, the Benson Figure Copy and Recall tests

Table 1. Demographic characteristics of the subsamples

	Development sample	Validation sample	*t* or χ^2, *p*
Age (*M*, *SD*)	69.98, 7.69	70.28, 7.82	−0.71, .479
Education (*M*, *SD*)	16.50, 2.43	16.41, 2.48	0.60, .547
% female	64.60%	63.00%	0.36, .551
% white	76.60%	81.00%	3.79, .052

(Possin, Laluz, Alcantar, Miller, & Kramer, 2011), and a phonemic fluency measure; the removal of one test (a digit-symbol coding task); and the replacement of two previous tests with royalty free versions—the Boston Naming Test (Goodglass, Kaplan, & Barresi, 2000) was replaced by the Multilingual Naming Test (MINT; Ivanova, Salmon, & Gollan, 2013) and the Wechsler Digit Span subtest was replaced with a comparable Number Span Test. Given these revisions, and the central role of the UDS3NB to large-scale aging and cognition projects, understanding and establishing means of interpreting change over time with this battery is of critical importance. Furthermore, these methods of assessing change should be accessible to clinicians and researchers in a user-friendly format that allows for accurate, straightforward interpretation of serial test data (Gavett et al., 2015). Thus, SRB methods need to be applied to the UDS3NB to validate these change scores as measures of transitional cognitive decline, and tools need to be developed to easily calculate change metrics with individual patients.

Current Study

In summary, the current study had four main goals: derive 6-month to 2-year SRB change indices for the UDS3NB in a sample classified as cognitively normal across baseline and first follow-up observations; assess the base rates of significant change at different *SD* cut-points; complete preliminary convergent and concurrent validity analyses of the SRB indices; and develop a user-friendly Excel calculator for deriving SRB change indices for individual patients.

Methods

Sample

We requested all available UDS data on June 11, 2019 collected at 39 Alzheimer's Disease Research Centers. The process of selecting cases for final analyses is summarized in Fig. 1. Because a number of the measures included in the UDS3NB require proficiency in English, the sample was limited to primary language English speakers. Next, because we were interested in calculating SRB indices using the UDS3NB, the sample was restricted to individuals who had received the UDS 3.0 at their initial visit. We then excluded individuals with diagnosed cognitive impairment at their baseline or time 2 follow-up visit. Determination of cognitive impairment was based on the Clinical Dementia Rating® (CDR) Dementia Staging Instrument[1] (Morris, 1993), a reliable and valid clinical interview for staging of dementia (Fillenbaum, Peterson, & Morris, 1996; Morris, 1997). Individuals with CDR $>$ 0 at time 1 (baseline) or time 2 (1-year follow-up) were excluded from the study. Individuals with known amyloid biomarker status were also removed to reflect the current state of affairs in most clinics. Next, because we were primarily interested in measuring performance among older adults, individuals below age 50 were excluded from the analyses. Finally, due to the fact that follow-up times could vary, we limited the sample to individuals with follow-up data from 6 months to 2 years from their baseline evaluations ($M_{\text{days}} = 423.32$, $SD = 84.18$). This step enabled the inclusion of a wide range of follow-up periods that might be used in research and clinical practice while limiting the influence of outliers. This final sample included participants with baseline data collected from March 2015 through July 2018. The sample was divided into two randomly selected subsets to form separate groups for development ($n = 627$) and validation ($n = 714$) of SRB indices. Note that because a truly random sampling method was used, sample sizes could be unequal. Demographic information for each subsample is presented in Table 1. There were no significant differences in the demographic makeup of the subsamples.

Measures

Cognitive tests. The cognitive tests in the UDS3NB have been described in detail in previous publications (Besser et al., 2018; Weintraub et al., 2018). We used a subset of these measures for our analyses. These included: the Montreal Cognitive Assessment

[1] Note that staging based on the UDS consensus diagnosis was not used to avoid criterion contamination, as rating clinicians utilize test data to make diagnostic determinations in the UDS.

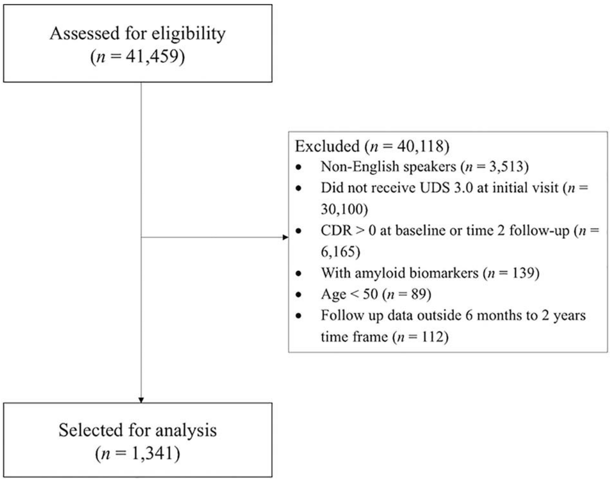
Assessed for eligibility
(n = 41,459)
Excluded (n = 40,118)
Non-English speakers (n = 3,513)
Did not receive UDS 3.0 at initial visit (n = 30,100)
CDR > 0 at baseline or time 2 follow-up (n = 6,165)
With amyloid biomarkers (n = 139)
Age < 50 (n = 89)
Follow up data outside 6 months to 2 years time frame (n = 112)
Selected for analysis
(n = 1,341)

Fig. 1. Sample selection flow chart.

(MoCA), a brief screener for cognitive impairment (Nasreddine et al., 2005); tests of simple number sequencing and letter-number sequencing, Trailmaking Test Parts A and B[2] (Partington & Leiter, 1949); the Craft Story (Craft et al., 1996), a test of immediate and delayed recall of verbal story information (though both thematic and verbatim recall scores are indexed, verbatim recall points were used in the current analyses); a measure of visuoconstruction and visual recall, the Benson Figure, (Possin et al., 2011); semantic (animals and vegetables) and letter (F- and L-words) verbal fluency tasks; number span forward and backward (total score used for the current study); and a confrontation naming measure, the Multilingual Naming Test (Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012; Ivanova et al., 2013). Descriptive statistics and test–retest correlations for each cognitive test in the normative sample are summarized in Table 2.

Clinical dementia rating scale. In addition to a categorical global score, the CDR includes a continuous measure of impairment based on clinician rating of patient functioning in six domains. They include memory, orientation, judgement and problem solving, community affairs, home and hobbies, and personal care. Domains are scored according to the level of impairment, including none (0), questionable (0.5), mild (1), moderate (2), and severe (3). Scores from each domain are summed to form an overall impairment score (0–18 point scale), known as the sum of box scores (CDR-SB). This score has excellent evidence for validity as an index of impairment in aging populations (Lynch et al., 2006; O'Bryant et al., 2008, 2010; Samtani, Raghavan, Novak, Nandy, & Narayan, 2014), with interrater and internal consistency reliability estimates in the acceptable to excellent range (Burke et al., 1988; Cedarbaum et al., 2013).

Analyses

SRB change index calculation. SRBs were calculated based on raw scores[3] using a multivariate version of the method outlined by McSweeny and colleagues (1993), with cases including missing data excluded list-wise (Hinton-Bayre, 2010). Specifically, using the development sample, each time 2 neuropsychological test score was regressed onto its corresponding time 1 neuropsychological test score, as well as demographic factors (age, sex, education, and race [coded as white vs. nonwhite]) and time (days since the baseline evaluation to account for minor variations in follow-up length). All predictors were entered as a block, conforming to previous publications with the UDS 3.0 that included demographic corrections (Devora et al., 2019; Kiselica et al., 2020a, 2020b; Weintraub et al., 2018). These results were then used to calculate a regression-based predicted score for the time 2 cognitive test. Finally, the SRB change index was calculated by subtracting the predicted time 2 test score, T'_2, from the observed time 2 test score, T_2, and then dividing this value by the standard error of the estimate (SEE), using the

[2] These measures were reverse coded for SRB analyses, such that higher scores were indicative of improved scores, in order to conform with scaling of other measures.

[3] Standardized scores can be used and will often yield equivalent findings to raw scores; however, some research suggests that raw score models better capture changes in memory over time than do standardized scores (Durant, Duff, & Miller, 2019). Using raw scores is also mathematically simpler, as it eliminates the need to convert to standard scores. Consequently, raw scores were used for all analyses.

Table 2. Mean change and test–retest correlations for raw cognitive variables across the baseline and first follow-up evaluations in the development subsample

Validation sample	Time 1 mean (*SD*)	Time 2 mean (*SD*)	t, p value	d_{rm}
MoCA	26.38 (2.56)	26.53 (2.63)	1.62, .105	0.06
Benson copy	15.58 (1.40)	15.49 (1.32)	−1.28, .202	−0.07
Benson recall	11.46 (2.92)	11.66 (2.92)	1.83, .068	0.07
Animal naming	21.79 (5.53)	21.43 (5.61)	−1.97, .049	−0.06
Vegetable naming	15.15 (3.91)	15.12 (3.99)	−0.19, .851	−0.01
Trailmaking part A	31.01 (11.57)	30.39 (11.47)	−1.49, .136	−0.05
Trailmaking part B	78.77 (37.18)	81.06 (42.79)	1.59, .114	0.06
Letter fluency	28.78 (8.12)	29.68 (8.46)	3.75, <.001	0.11
Craft story IR	21.92 (6.27)	22.50 (6.24)	2.35, .019	0.09
Craft story DR	19.15 (6.29)	19.74 (6.55)	2.40, .017	0.09
Numbers forward	8.46 (2.39)	8.39 (2.38)	−0.91, .365	−0.03
Numbers backward	7.26 (2.17)	7.28 (2.24)	0.18, .854	0.01
MINT	30.23 (1.87)	30.39 (1.94)	3.21, .001	0.08
Cognitive variable	Overall r, p ($n = 627$)	6–12 months r, p ($n = 150$)	13–18 months r, p ($n = 430$)	19–24 months r, p ($n = 47$)
MoCA	.63, <.001	.71, <.001	.62, <.001	.62, <.001
Benson copy	.31, <.001	.31, <.001	.34, <.001	.04, .803
Benson recall	.58, <.001	.58, <.001	.58, <.001	.58, <.001
Animal naming	.66, <.001	.65, <.001	.69, <.001	.72, <.001
Vegetable naming	.57, <.001	.48, <.001	.57, <.001	.73, <.001
Trailmaking part A	.61, <.001	.74, <.001	.57, <.001	.62, <.001
Trailmaking part B	.61, <.001	.64, <.001	.60, <.001	.81, <.001
Letter fluency	.75, <.001	.71, <.001	.78, <.001	.57, <.001
Craft story IR	.55, <.001	.44, <.001	.58, <.001	.52, <.001
Craft story DR	.56, <.001	.39, <.001	.62, <.001	.56, <.001
Numbers forward	.67, <.001	.66, <.001	.68, <.001	.57, <.001
Numbers backward	.60, <.001	.58, <.001	.59, <.001	.69, <.001
MINT	.78, <.001	.83, <.001	.76, <.001	.79, <.001

Notes: d_{rm} = Cohen's *d* for repeated measures; MoCA = Montreal Cognitive Assessment total score; IR = immediate recall; DR = delayed recall; numbers forward = number span forward; numbers backward = number span backward; MINT = multilingual naming test.

following formula:SRB Index $= (T_2 - T_2')/\mathrm{SEE}$. This process yields a regression-based *z*-score for change that accounts for test–retest reliability, practice effects, regression to the mean, measurement error, and demographic effects.

Of note, an alternative method of creating SRB change scores is available, which differs from the McSweeny and colleagues (1993) approach in how the standard error term is calculated (Crawford & Howell, 1998). The McSweeny approach was chosen for two reasons. First, it is more widely used in neuropsychology and may be more easily understood by practicing neuropsychologists (Hammers & Duff, 2019; Hinton-Bayre, 2010). Second, it is less computationally intensive (Crawford & Howell, 1998; Hammers & Duff, 2019), which was important for our goal of creating an easy-to-use calculator for deriving SRB indices. Importantly, the results from these two approaches have been found to be closely comparable, particularly with large, heterogeneous samples, like the one used in the current study (Crawford & Howell, 1998; Hammers & Duff, 2019; Hinton-Bayre, 2010).

Base rates of change at different cut-points. SRB formulas derived from the development sample were then applied to create SRB indices in the validation sample. SRB indices are expressed as *z*-score values, such that different *SD* cut-points can be set for assessing whether a change should be considered statistically unlikely. The most common cut-point used is ±1.645 *SD*. This cut-point ensures that only 10% of scores will be identified as abnormal (i.e., in the bottom or top 5% of the normal distribution for change) and is admittedly arbitrary (Duff, 2012; Hill, 2019). Consequently, we present base rates of objective decline not only at the −1.645 cut-point but also at the −1.96 and −1.282 cut-offs (corresponding to 5% and 20% of scores being identified as abnormal, respectively). Base rates of low scores are reported for each measure individually, as well when considering the whole battery.

Preliminary SRB index validation. Convergent validity of SRB indices was assessed by examining correlations among resultant *z*-scores from the respective cognitive variables. Small-to-moderate positive associations were expected between methodologically and theoretically related variables (e.g., Trails A with Trails B; animal fluency with vegetable fluency). Second, we examined concurrent validity of each SRB index by assessing correlations between SRB indices and measure of functional abilities at time 2 (CDR-SB). Given that the sample included presumably cognitively normal individuals, restriction in range for functional measures was anticipated, such that only small negative associations between SRB indices and CDR-SB were expected. We also examined the relationship between the number of significant test decreases and CDR-SB and hypothesized that there would be small but significant positive associations. The alpha level was set at 0.05 and no corrections for multiple tests were used, given the preliminary nature of the validation analyses. Due to the expectation that variables would not be normally distributed, we report Kendall's rank correlation coefficients for these analyses.

SRB tool development. SRB formulas from the development sample were transported to an Excel spreadsheet. The sheet was setup to allow an individual's raw scores to be entered and their SRB index to be automatically calculated. SRB index values were separately expressed in terms of percentile ranks and categorical cut-offs for statistically unlikely change (see Results section for details on the chosen cut-offs).

Results

Descriptive Statistics

Table 2 presents descriptive statistics for cognitive variables in the normative sample across time 1 and time 2. There were significant mean level increases in test scores from time 1 to time 2 on letter fluency, the Craft Story, and the MINT. There was a significant mean level decrease for animal fluency score from time 1 to time 2. Remaining mean level test score differences across time was nonsignificant. Effect sizes for significant tests were in the negligible to very small range.

Overall, test–retest correlations ranged from .31 (Benson Figure Copy) to .78 (MINT). As expected, correlations tended to be similar when examined at adjacent measurement points (6–12, 12–18, and 19–24 month follow-up periods). The exception was the Benson Copy, for which there was a nonsignificant relationship in scores across assessment points in the group reassessed at 19–24 months.

Regression Results in the Development Subsample

Results of the regression analyses predicting time 2 cognitive test scores from time 1 test scores, demographic information, and time elapsed between evaluations are presented in Table 3. Additional descriptive information on observed, predicted, and SRB z-scores is available in Supplementary Table 1. R^2 values ranged from moderate to very large (.14–.62). Baseline test scores were strong and consistent positive predictors of time 2 test scores. In contrast, older age and lower education were consistently associated with reduced scores at time 2. Effects of sex were mixed. Women performed better at time 2 on vegetable fluency, whereas men had better scores on the Benson and number span tasks. Race was not a significant predictor of time 2 test scores. Time elapsed tended not to be significantly related to time 2 test scores, though for Trailmaking Part B and the Benson Copy; greater time between evaluations was associated with reduced test scores. Parameters from these regression analyses were used to create formulas to calculate SRB change indices for each cognitive variable in the validation sample.

Base Rates of Change at Different Cut-Points in the Validation Subsample

In the validation subsample, we assessed base rates (i.e., percentage of individuals falling above/below a given cut-point) of change on individual tests (see Fig. 2). As might be expected, base rates of significant decreases on single measures corresponded fairly well with the cut-off chosen to define a significant change.

In addition to examining base rates of significant change for individual tests, we examined base rates of change across the entire UDS3NB battery. That is, we calculated the likelihood of having one, two, or three or more significant changes in test scores across the UDS3NB battery at the three score cut-offs: −1.645, −1.96, and −1.282. The results are presented graphically in Fig. 2. As expected, the likelihood of a significant change differed as a function of the chosen cut-off: Stricter cut-offs lead to a lower likelihood of having a score difference that qualified as statistically unlikely. Having at least one significant test score change was fairly common (26.70%–58.10%), whereas having two or three test scores change significantly was increasingly rare (1.20%–11.10%).

Table 3. Results of linear regressions in the development subsample

	Intercept	Age Coefficient	Sex Coefficient	Education Coefficient	Race Coefficient	Time Coefficient	Baseline score Coefficient	R^2	SEE
MoCA	10.66***	−0.02	0.18	0.12**	−0.05	−0.002	0.60***	.42***	2.01
Benson copy	13.67***	−0.03***	−0.25*	0.01	0.01	−0.001*	0.29***	.14***	1.24
Benson recall	9.00***	−0.05***	−0.45*	0.03	−0.02	0.000	0.54***	.36***	2.35
Animal fluency	4.89*	−0.03	0.08	0.24**	−0.01	0.001	0.65***	.47***	4.05
Vegetable fluency	6.02**	−0.04*	0.87**	0.17**	−0.06	−0.001	0.53***	.34***	3.26
Trailmaking part A	-0.66	−0.19***	−0.52	0.24	−0.15	−0.004	0.57***	.38***	9.02
Trailmaking part B	-2.69	−0.57**	−1.07	1.72**	−0.39	−0.04*	0.62***	.37***	33.40
Letter fluency	8.70**	−0.04	−0.25	0.25*	−0.13	−0.003	0.76***	.57***	5.59
Craft story IR	9.31**	−0.01	0.10	0.09	−0.04	0.001	0.54***	.30***	5.24
Craft story DR	9.94**	−0.02	−0.07	0.01	−0.02	−0.001	0.59***	.32***	5.44
Numbers forward	4.94***	−0.02	−0.36*	0.01	−0.02	−0.001	0.65***	.46***	1.75
Numbers backward	3.95***	−0.02	−0.47**	0.05	−0.03	0.001	0.59***	.37***	1.79
MINT	7.50***	−0.02*	−0.09	0.04*	0.02	−0.001	0.79***	.62***	1.20

Notes: Male coded as 1, female coded as 2. White coded as 1, non-white coded as 2. Time expressed as days between T1 (baseline) and T2 (first follow up). SEE = standard error of the estimate, MoCA = Montreal Cognitive Assessment total score; IR = immediate recall; DR = delayed recall; numbers forward = number span forward; numbers backward = number span backward; MINT = multilingual naming test.

*$p < .05$.

**$p < .01$.

***$p < .001$.

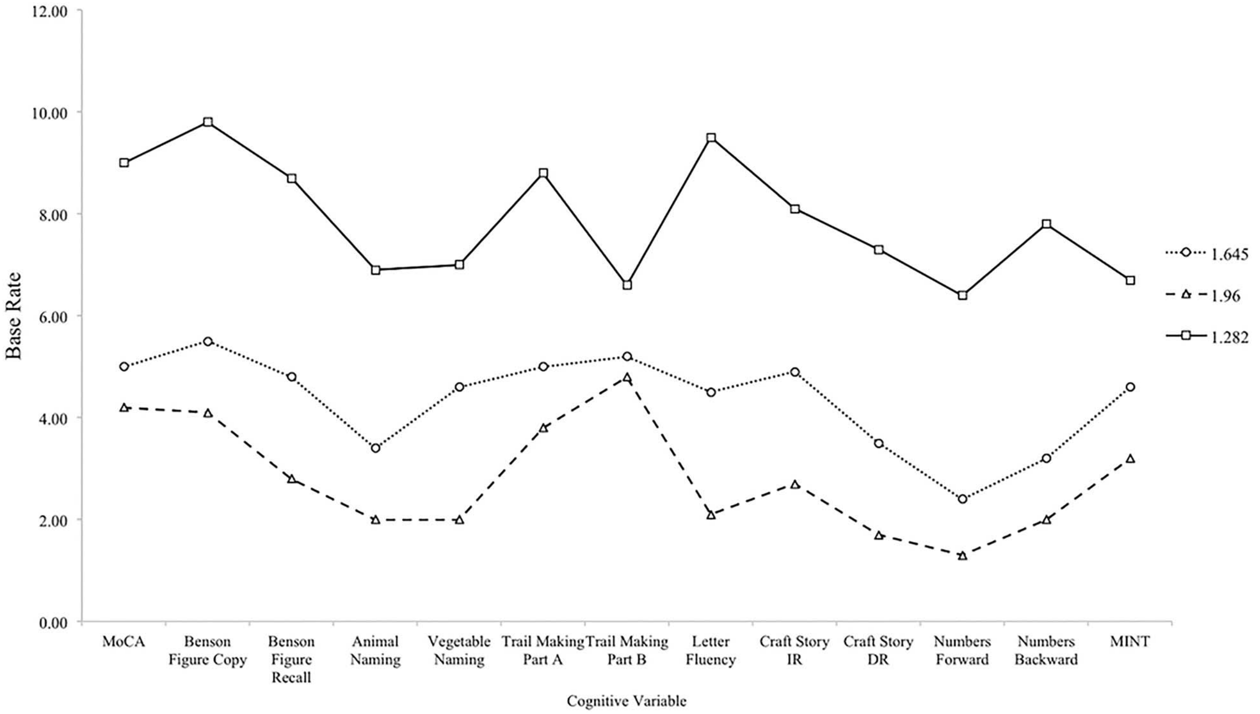

Fig. 2. Base rate of standardized regression-based decline at different cut-offs for each cognitive test in the validation subsample.

Convergent Validity of SRB Indices

We examined the convergent associations of SRB indices with each other in the validation subsample (see Table 4). As a fairly global measure of cognitive abilities, MoCA SRB scores were significantly, albeit weakly, positively correlated with most other SRB scores. Significant correlations between other SRB indices tended to reflect measure similarity. Indeed, positive correlations were observed between figure copy and recall SRB scores ($r = .23$, $p < .001$), semantic fluency SRB scores ($r = .21$,

Table 4. Convergent correlations of standardized regression-based (SRB) indices and concurrent correlations of SRB indices with CDR-SB

	2	3	4	5	6	7	8	9	10	11	12	13	14
MoCA	.11**	.15***	.05	.13**	.15***	.15***	.14***	.14***	.18***	.13**	.11**	.13**	-.08**
Benson copy		.23***	.07	.20	-.04	-.01	.05	.01	.02	.04	.02	.02	-.02
Benson recall			.14***	.13**	.02	.04	.04	.06	.06	.11**	.12**	.16**	.06
Animal fluency				.21**	.03	.03	.12**	.03	.02	.02	.12**	.06	-.06
Vegetable fluency					.05	.10*	.05	.10*	.07	.07	.09*	.12**	-.05
Trailmaking part A						.25***	-.02	.03	.05	.07	.10*	.06	-.11**
Trailmaking part B							.07	.09*	.10*	.05	.06	.08*	-.05
Letter fluency								.09*	.08*	.01	.07	.01	-.03
Craft story IR									.75***	.03	.08*	.07	-.03
Craft story DR										.01	.05	.10*	-.05
Numbers forward											.22***	.03	-.04
Numbers backward												.07	-.10**
MINT													.03
Time 2 CDR-SB													

Notes: Correlations between SRB scores and CDR-SB scores reflect Kendall's rank correlations. MoCA = Montreal Cognitive Assessment total score; IR = immediate recall; DR = delayed recall; numbers forward = number span forward; numbers backward = number span backward; MINT = multilingual naming test; CDR-SB = clinical dementia rating sum of box scores.

$^{*}p < .05$.

$^{**}p < .01$.

$^{***}p < .001$.

$p < .001$), trailmaking SRB scores ($r = .25$, $p < .001$), number span SRB scores ($r = .22$, $p < .001$), and story memory and recall SRB scores ($r = .75$, $p < .001$).

Concurrent Validity of SRB Indices

Trailmaking Part A and Number Span Backwards SRB scores were significantly negatively correlated with CDR-SB. Additionally, having an increasing number of significant test decreases (i.e., a higher number of SRB change scores at or below the cut-point) was associated with a higher CDR-SB score for all cut-offs (≤ -1.96: $\tau = .10$, $p = .019$; ≤ -1.645: $\tau = .10$, $p = .044$; and ≤ -1.282: $\tau = .10$, $p = .004$).

SRB Tool Creation

These formulas were placed into an Excel sheet. This sheet enables an automatic calculation of SRB change indices upon entry of demographic data and raw test scores. It also generates qualitative descriptors at different *z*-score cut-points for change, including ±1.645, ±1.96, and ±1.282. Specifically, these columns indicate whether an individual has experienced a significant increase in performance, a significant decrease in performance, or no significant change. The calculator is available in the online supplementary material.

Discussion

Change and Reliability of the UDS3NB

The psychometric properties of the UDS3NB frame, the discussion of an evaluation of change on this measure. To that end, our current results revealed that mean level changes over test–retest intervals of 6–24 months in our sample of cognitively normal older adults were typically nonsignificant or negligible in effect size, such that they are not likely to be practically meaningful. Consistently, research has shown that in the absence of neuropathology, cognitive declines over time in older adults tend to be slow (Harada, Love, & Triebel, 2013). These results further suggest relatively limited practice effects on UDS3NB measures, which is inconsistent with findings from a meta-analysis by Calamia, Markon, and Tranel (2012). The authors reported that practice effects are more common when shorter test–retest intervals are used and when the sample is younger. Thus, the use of an older sample and longer follow-up periods may explain our lack of findings. Alternatively, it may be that practice effects do not emerge for the UDS3NB over only two assessments. Indeed, such a pattern was reported by Gavett and colleagues (2015) in their analyses of the UDS 2.0, wherein practice effects only emerged after the assessment was repeated more than once.

Of note, despite a lack of mean-level practice effects, there were many participants who demonstrated improved scores from time 1 to time 2 at the individual level. Consequently, research on factors that influence whether practice effects occur and the clinical implications of the presence or absence of practice effects is important (see Duff et al., 2017).

In addition to mean-level change, we examined test–retest correlations for the UDS3NB measures. Unfortunately, test–retest reliability for common neuropsychological measures is not always reported; when they are reported, these analyses are often completed on small samples. Moreover, researchers often utilize follow-up periods that do not mimic the realities of clinical practice, wherein testing is rarely repeated until at least 1 year has passed (Lezak et al., 2012). For example, the Neuropsychological Assessment Battery (NAB) manual (White & Stern, 2003) tabulates uncorrected test–retest correlations ranging from .08 to .84 among a sample of 37 older adults tested twice over a 6-month time frame. Thus, the current test–retest analyses, which were completed on a sample of over 1,000 participants tested over a 6-month to 2-year timeframe, represent a substantial improvement on those noted in previous work.

Against that backdrop, test–retest correlations in our sample ranged from .31 to .78. While test–retest correlations closer to 1.0 are preferable, high stability is not necessarily expected with the measurement of psychological constructs, such as cognitive performance, which are impacted by a host of factors over time and constrained by measure reliability properties (Lezak et al., 2012; Miller & Lovler, 2018). That being said, the values obtained from the UDS3NB were roughly comparable to those presented for published batteries, such as the NAB, as well as meta-analytic findings (Calamia et al., 2013), suggesting that the UDS3NB demonstrates at least comparable reliability to other established sets of tests.

Of note, one test on the UDS3NB (the Benson Figure copy) demonstrated a fairly low test–retest correlation overall (.31), especially in the group of individuals who were retested after 1.5 years (.04). Our findings are in line with previous work demonstrating comparatively low stability for the Rey Complex Figure Copy (Calamia et al., 2013; Meyers & Meyers, 1995). Additionally, the observation of very low reliability after long follow-up conforms with findings for the NAB figure copy test (estimated at .08 in the manual). One potential explanation for the reduced stability of figure copy tasks is that they rely on somewhat subjective scoring criteria. Additionally, these tests tend to be scored on scales with small ranges, such that they are highly impacted by minor shifts in score (e.g., a three-point change on a 40-point scale is unlikely to dramatically impact rank order, whereas a three-point change on a 17-point scale will influence rank order to a high degree).

Factors that Impacted Change in Cognitive Test Scores

Next, in the course of calculating SRB indices, we examined a number of predictors of time 2 test scores. Individuals with higher levels of education and higher baseline test scores tended to have stronger time 2 cognitive scores. This research is consistent with previous research, demonstrating superior scores over time for groups with higher premorbid abilities, and supports the cognitive reserve hypothesis (Lenehan, Summers, Saunders, Summers, & Vickers, 2015; Rapport, Brines, Theisen, & Axelrod, 1997; Stern, 2002). Additionally, as would be expected based on past studies (Salthouse, 2010), younger individuals tended to perform better on cognitive testing at time 2. The influence of sex on time 2 test scores was variable. While the finding of higher scores among men on visual tasks is consistent with prior research in older adults, though their improved number span scores were surprising (Munro et al., 2012). Additionally, women performed better on vegetable fluency, replicating prior cross-sectional findings with the UDS that likely reflect differences in socialization and social roles (Kiselica et al., 2020b). Time between evaluations did not typically appear to influence test scores. Similarly, race was not a significant predictor of time 2 test scores. This finding is consistent with recent research, which suggest that race influences baseline test score but not change over time (Gross et al., 2015).

SRB Change Scores and Clinically Meaningful Change

Using the regression parameters, we created SRB change indices using a multivariate version of the method outline by McSweeny and colleagues (1993). After calculating SRB change scores, we examined base rates of significant decreases to help differentiate expected patterns of change from those possibly suggestive of pathological decline (Hill, 2019). SRB change scores for individual tests tended to fall at the tails of the distribution at rates expected based on the cut-off chosen to define a significant change. For instance, about 5% of individuals would be expected to have a significant decrease at the −1.645 or below cut-point and observed values range from about 3% to 6% (see Fig. 1). These findings suggest that in contexts where only a single test has been administered (an exceedingly rare case in actual practice), an SRB change at a given cut-off can be interpreted as both statistically and clinically meaningful.

The same cannot be said when examining the patterns of change scores across the entire UDS3NB. Indeed, having at least one test score decrease significantly was fairly common, occurring in 26.70%–58.10% of participants, depending on the chosen *SD* criterion (see Fig. 2). Rates of having multiple test scores change were much lower. For instance, only 1.20%–11.10% of individuals had a significant test score decrease on three or more tests. These findings are consistent with those of Brooks and colleagues (2016), who reported high rates (up to 39%) of one or more significant reliable change scores for the NAB memory module and the Wechsler Memory Scale-IV among older adults. Taken together, existing research suggests that having one statistically significant change score in a battery of tests is not clinically meaningful; rather, it is expected even for cognitively intact individuals. In our sample, it was even fairly common (14.40% of participants) to have a significant decrease on two tests at the typically used ≤ -1.645 *SD* cut-point. Thus, it is likely that if one uses the ≤ -1.645 *SD* criterion, he/she should only consider a decline pattern involving three or more tests (occurring in 4.30% of participants in our normative sample) as clinically meaningful on the UDS3NB.

There are several caveats to interpreting these findings. First, it must be acknowledged that the likelihood of obtaining one or more significant test score changes depends on the number of tests administered (Brooks, Iverson, & White, 2009a; Oltra-Cucarella, Sánchez-SanSegundo, Rubio-Aparicio, Arango-Lasprilla, & Ferrer-Cascales, 2019). Thus, our base rate findings will be most applicable to interpreting the specific set of UDS3NB tests used in the current paper. Second, we examined the base rates of change across the entire UDS3NB, and it is possible that more clinically meaningful patterns could emerge by examining within-domain patterns of change or discrepancies in change between different tests (Devora et al., 2019; Jak et al., 2016; Litvan et al., 2012). Similarly, some researchers have put forth methods to examine change across a whole battery (Woods et al., 2006), and it may be informative to examine change in summary or factor scores over time (Kiselica et al., 2020b). Future research should explore the base rates of change in other test batteries, as well as alternative methods of interpreting SRB change. To support such efforts, an Excel calculator is provided in the online supplementary material that allows for easy derivation of SRB change indices from raw UDS3NB data.

Preliminary Convergent and Concurrent Validity of SRB Indices

SRB indices calculated in the validation sample demonstrated positive correlations with similar measures. Results closely mirrored relationships suggested by a recent factor analysis of the UDS3NB (Kiselica et al., 2020b); that is, associations between SRB scores were observed between processing speed/executive measures (Trailmaking Parts A and B), attention (Number Span Backwards and Forwards), visual (Benson Figure copy and recall), memory (Craft Story immediate and delayed recall), and language measures (vegetable and animal fluency). Most relationships were small-to-moderate in effect size (as opposed to moderate-to-large associations reported in cross-sectional studies), likely because change scores tend to be less reliable than single time-point scores (Allison, 1990). However, the relationship between scores on the Craft Story test was quite strong ($r = .75$), likely reflecting then dependency of the delayed recall score on how much is initially learned.

We also examined concurrent associations between SRB scores and CDR-SB, finding some modest but significant results. Two factors played a role in limiting the size of effects. First, as anticipated, there was range restriction in CDR-SB scores. Indeed, as a result of our selection criteria, CDR-SB scores ranged from 0 to 1 (687 individuals with a score of 0, 26 with a score of 0.5, and one with a score of 1). Second, much of the predictive variance is partialed out of SRB scores; that is, baseline cognitive performance, time, and demographic factors are taken into account in their calculation. Thus, the analyses are very similar to a hierarchical regression and assess the extent to which change scores are associated with CDR-SB after controlling for several important factors.

Results indicated that Trailmaking Part A and Number Span Backwards SRB scores were significantly negatively associated with CDR-SB. This finding is unsurprising, given the primacy of processing speed and executive functions in mediating functional skills (Bezdicek, Stepankova, Novakova, & Kopecek, 2016; Martyr & Clare, 2012; Royall & Palmer, 2014). Additionally, having an increasing number of tests change significantly was associated with a higher CDR-SB, providing preliminary support for the base rate approach to interpreting SRB data (Brooks et al., 2016). That is, findings suggest that the number of significantly changed scores is a clinically important variable to consider, in addition to examining changes in particular test scores over time. In particular, our analyses suggest that having three or more significant change scores across the UDS3NB is rare in a normal sample and may be indicative of atypical performance (see Fig. 3). Of course, further longitudinal research on the number of significant change scores that predicts conversion to clinical states is necessary to confirm this assumption.

Implications of the Current Research

There is a growing body of literature that suggests that mild cognitive declines presage the accumulation of Alzheimer's disease biomarkers and the future development of MCI and dementia (Edmonds et al., 2015; Thomas et al., 2018, 2019). Our findings mirror these results and extend them, as this is the first research to our knowledge that supports a cross-sectional connection between transitional cognitive declines and extremely subtle changes in day-to-day functioning. Stated another way,

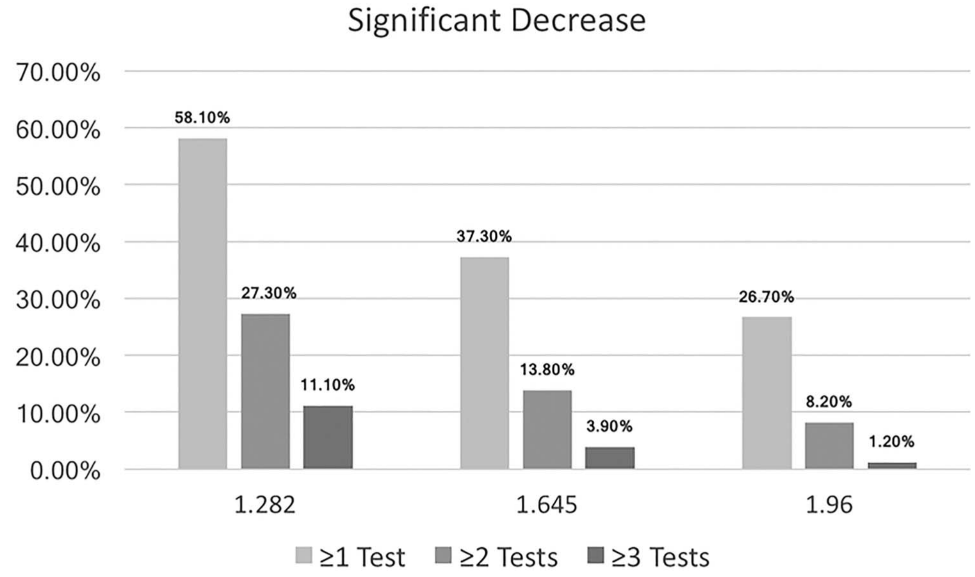

Fig. 3. Base rates for the numbers of tests with significant standardized regression-based decreases in the validation subsample for the Uniform Data Set 3.0 Neuropsychological Battery at different cut-off points.

prior work has suggested that transitional cognitive declines predict future functional impairment, whereas our work indicates that even very mild forms of cognitive decline are likely to have a very subtle impact on daily functioning at their onset.

This research also has implications for recently proposed Alzheimer's disease research framework from NIA-AA (Jack et al., 2018). Under this framework, there will be a biomarker-defined Alzheimer's Continuum, including two preclinical disease stages (i.e., prior to the development of MCI). Stage 1 will be defined by "no evidence of recent cognitive decline" (p. 55) from cognitive test data, whereas stage 2 will be indicated by the presence "evidence of subtle decline on longitudinal cognitive testing" (p. 55). The current SRB methodology may provide a means of indexing the presence/absence of such subtle declines on longitudinal testing, though important questions remain. For instance, what is the optimal cut-point to define transitional cognitive decline? And second, how many test scores need to decline to be considered meaningful? Future work can answer these questions by exploring the sensitivity and specificity of different methods for identifying those who progress to MCI from those who do not. Such research may fruitfully include machine learning techniques to identify optimal cut-offs and indicative patterns of change over time (Lin et al., 2018). Alternatively, identifying a particular cut-point or number of significant change scores may be less important than examining the slope/shape of cognitive test score change over time (Papp et al., 2019).

Limitations

There were some limitations of the current investigation worth noting. First, the use of the UDS sample comes with certain shortcomings. Certainly, data in the UDS comes primarily from research participants at Alzheimer's Disease Research Centers. Thus, while results are likely highly applicable in research settings, they may be less applicable in other contexts (e.g., when working in a purely clinical setting). Furthermore, the UDS sample consists of primarily college-educated, white individuals, such that findings may not be applicable to individuals from other backgrounds. Efforts to diversify the UDS sample are ongoing, such that future research will be able to replicate the current research using data from lower educated and more racially inclusive groups. Second, the SRB indices calculated in the current analyses only covered a 6-month to 2-year time frame; therefore, test changes over longer time intervals cannot be interpreted using our calculator. Future studies could calculate SRB change values using longer follow-ups. Third, the validation analyses were conducted in a sample of putatively cognitively normal individuals, limiting the ability to detect relationships of SRB indices with functional declines. Subsequent studies will validate the SRB indices in clinical samples (e.g., those with MCI or dementia).

Conclusions

Examining changes in cognitive test scores over time is of increasing importance in the current landscape of aging and Alzheimer's disease research. SRB change indices can help distinguish expected cognitive changes from those indicative of the presence of pathology. This paper provides the preliminary validation of SRB indices, though more work is needed in clinical populations. To support such research, we present a user-friendly tool for calculating SRB values as well as base rate data to aid in their interpretation.

Supplementary material

Supplementary material is available at *Archives of Clinical Neuropsychology* online.

Acknowledgements

The NACC database is funded by NIA/NIH (Grant U01 AG016976). NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Funding

This work was supported by an Alzheimer's Association Research Fellowship (2019-AARF-641693, PI Andrew Kiselica, PhD) and the 2019–2020 National Academy of Neuropsychology Clinical Research Grant (PI Andrew Kiselica, PhD).

Conflict of Interest

None declared.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

Andrews, J. S., Desai, U., Kirson, N. Y., Zichlin, M. L., Ball, D. E., & Matthews, B. R. (2019). Disease severity and minimal clinically important differences in clinical outcome assessments for Alzheimer's disease clinical trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *5*, 354–363.

Besser, L., Kukull, W., Knopman, D. S., Chui, H., Galasko, D., Weintraub, S., et al. (2018). Version 3 of the National Alzheimer's coordinating Center's uniform data set. *Alzheimer Disease and Associated Disorders*, *32*(*4*), 351.

Bezdicek, O., Stepankova, H., Novakova, L. M., & Kopecek, M. (2016). Toward the processing speed theory of activities of daily living in healthy aging: Normative data of the functional activities questionnaire. *Aging Clinical and Experimental Research*, *28*(*2*), 239–247.

Brooks, B. L., Holdnack, J. A., & Iverson, G. L. (2016). To change is human: "abnormal" reliable change memory scores are common in healthy adults and older adults. *Archives of Clinical Neuropsychology*, *31*(*8*), 1026–1036. doi: 10.1093/arclin/acw079.

Brooks, B. L., Iverson, G. L., & White, T. (2009a). Advanced interpretation of the neuropsychological assessment battery with older adults: Base rate analyses, discrepancy scores, and interpreting change dagger. *Archives of Clinical Neuropsychology*, *24*(*7*), 647–657. doi: 10.1093/arclin/acp061.

Brooks, B. L., Sherman, E. M., Iverson, G. L., Slick, D. J., & Strauss, E. (2011). *Psychometric foundations for the interpretation of neuropsychological test results. The little black book of neuropsychology* (, pp. 893–922). New York, NY: Springer.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009b). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, *50*(*3*), 196–209. doi: 10.1037/a0016066.

Burke, W. J., Miller, J. P., Rubin, E. H., Morris, J. C., Coben, L. A., Duchek, J., et al. (1988). Reliability of the Washington University clinical dementia rating. *Archives of Neurology*, *45*(*1*), 31–32.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(*4*), 543–570. doi: 10.1080/13854046.2012.680913.

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist*, *27*(*7*), 1077–1105.

Cedarbaum, J. M., Jaros, M., Hernandez, C., Coley, N., Andrieu, S., Grundman, M., et al. (2013). Rationale for use of the clinical dementia rating sum of boxes as a primary outcome measure for Alzheimer's disease clinical trials. *Alzheimer's & Dementia*, *9*(*1*), S45–S55.

Craft, S., Newcomer, J., Kanne, S., Dagogo-Jack, S., Cryer, P., Sheline, Y., et al. (1996). Memory improvement following induced hyperinsulinemia in Alzheimer's disease. *Neurobiology of Aging*, *17*(*1*), 123–130.

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, *20*(*5*), 755–762.

Cummings, J. (2019). The National Institute on Aging—Alzheimer's Association framework on Alzheimer's disease: Application to clinical trials. *Alzheimer's & Dementia*, *15*(*1*), 172–178.

Cummings, J., Lee, G., Ritter, A., Sabbagh, M., & Zhong, K. (2019). Alzheimer's disease drug development pipeline: 2019. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *5*, 272–293.

Devora, P. V., Beevers, S., Kiselica, A. M., & Benge, J. F. (2019). Normative data for derived measures and discrepancy scores for the uniform data set 3.0 neuropsychological battery. *Archives of Clinical Neuropsychology*. *35*(*1*), 75–89.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(*3*), 248–261.

Duff, K., Atkinson, T. J., Suhrie, K. R., Dalley, B. C. A., Schaefer, S. Y., & Hammers, D. B. (2017). Short-term practice effects in mild cognitive impairment: Evaluating different methods of change. *Journal of Clinical and Experimental Neuropsychology*, *39*(*4*), 396–407.

Duff, K., Suhrie, K. R., Dalley, B. C., Anderson, J. S., & Hoffman, J. M. (2019). External validation of change formulae in neuropsychology with neuroimaging biomarkers: A methodological recommendation and preliminary clinical data. *The Clinical Neuropsychologist*, *33*(*3*), 478–489.

Durant, J., Duff, K., & Miller, J. B. (2019). Regression-based formulas for predicting change in memory test scores in healthy older adults: Comparing use of raw versus standardized scores. *Journal of Clinical and Experimental Neuropsychology*, *41*(*5*), 460–468.

Edmonds, E. C., Delano-Wood, L., Galasko, D. R., Salmon, D. P., Bondi, M. W., & Alzheimer's Dis Neuroimaging, I (2015). Subtle cognitive decline and biomarker staging in preclinical Alzheimer's disease. *Journal of Alzheimers Disease*, *47*(*1*), 231–242. doi: 10.3233/jad-150128.

Fillenbaum, G., Peterson, B., & Morris, J. (1996). Estimating the validity of the clinical dementia rating scale: The CERAD experience. *Aging Clinical and Experimental Research*, *8*(*6*), 379–385.

Gavett, B. E., Ashendorf, L., & Gurnani, A. S. (2015). Reliable change on neuropsychological tests in the uniform data set. *Journal of the International Neuropsychological Society*, *21*(*7*), 558–567.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A multilingual naming test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*(*3*), 594–615.

Goodglass, H., Kaplan, E., & Barresi, B. (2000). *Boston Diagnostic Aphasia Examination Record Booklet*. Philadelphia, PA: Lippincott Williams & Wilkins.

Gross, A. L., Mungas, D. M., Crane, P. K., Gibbons, L. E., MacKay-Brandt, A., Manly, J. J., et al. (2015). Effects of education and race on cognitive decline: An integrative study of generalizability versus study-specific results. *Psychology and Aging*, *30*(*4*), 863.

Hammers, D. B., & Duff, K. (2019). Application of different standard error estimates in reliable change methods. *Archives of Clinical Neuropsychology*, acz054, doi: 10.1093/arclin/acz054.

Harada, C. N., Love, M. C. N., & Triebel, K. L. (2013). Normal cognitive aging. *Clinics in Geriatric Medicine*, *29*(*4*), 737–752.

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). 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. *The Clinical Neuropsychologist*, *24*(*8*), 1267–1278.

Hill, S. W. (2019). Components and methods of evaluating reliable change in cognitive function neurosurgical neuropsychology. In C. Pearson, E. Ecklund-Johnson, S. Gale (Eds.), *Neurosurgical Neuropsychology*, pp. 39–61. San Diego, CA: Elsevier.

Hinton-Bayre, A. D. (2010). Deriving reliable change statistics from test–retest normative data: Comparison of models and mathematical expressions. *Archives of Clinical Neuropsychology*, *25*(*3*), 244–256.

Hinton-Bayre, A. D. (2016). Clarifying discrepancies in responsiveness between reliable change indices. *Archives of Clinical Neuropsychology*, *31*(*7*), 754–768.

Ivanova, I., Salmon, D. P., & Gollan, T. H. (2013). The multilingual naming test in Alzheimer's disease: Clues to the origin of naming impairments. *Journal of the International Neuropsychological Society*, *19*(*3*), 272–283.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement*, *14*(*4*), 535–562. doi: 10.1016/j.jalz.2018.02.018.

Jak, A. J., Preis, S. R., Beiser, A. S., Seshadri, S., Wolf, P. A., Bondi, M. W., et al. (2016). Neuropsychological criteria for mild cognitive impairment and dementia risk in the Framingham heart study. *Journal of the International Neuropsychological Society*, *22*(*9*), 937–943.

Kiselica, A. M., Webber, T., & Benge, J. (2020a). Using multivariate base rates of low scores to understand early cognitive declines on the uniform data set 3.0 neuropsychological battery. *Neuropsychology*, e-publication ahead of print. doi: 10.1037/neu0000640.

Kiselica, A. M., Webber, T. A., & Benge, J. F. (2020b). The uniform data set 3.0 neuropsychological battery: Factor structure, invariance testing, and demographically-adjusted factor score calculation. *Journal of the International Neuropsychological Society*, *26*(*6*), 576–586. doi: 10.1017/S135561772000003X.

Kornak, J., Fields, J., Kremers, W., Farmer, S., Heuer, H. W., Forsberg, L., et al. (2019). Nonlinear Z-score modeling for improved detection of cognitive abnormality. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *11*(*C*), 797–808.

Lenehan, M. E., Summers, M. J., Saunders, N. L., Summers, J. J., & Vickers, J. C. (2015). Relationship between education and age-related cognitive decline: A review of recent research. *Psychogeriatrics*, *15*(*2*), 154–162.

Lezak, M., Howieson, D., & Loring, D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford.

Liew, T. M. (2019). Developing a brief neuropsychological battery for early diagnosis of cognitive impairment. *Journal of the American Medical Directors Association*, *20*(*8*), 1054 e1011-1054, e1020.

Lin, M., Gong, P., Yang, T., Ye, J., Albin, R. L., & Dodge, H. H. (2018). Big data analytical approaches to the NACC dataset: Aiding preclinical trial enrichment. *Alzheimer Disease and Associated Disorders*, *32*(*1*), 18.

Litvan, I., Goldman, J. G., Tröster, A. I., Schmand, B. A., Weintraub, D., Petersen, R. C., et al. (2012). Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society task force guidelines. *Movement Disorders*, *27*(*3*), 349–356.

Lynch, C., Walsh, C., Blanco, A., Moran, M., Coen, R., Walsh, J., et al. (2006). The clinical dementia rating sum of box score in mild dementia. *Dementia and Geriatric Cognitive Disorders*, *21*(*1*), 40–43.

Martyr, A., & Clare, L. (2012). Executive function and activities of daily living in Alzheimer's disease: A correlational meta-analysis. *Dementia and Geriatric Cognitive Disorders*, *33*(*2–3*), 189–203.

McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Lüders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist*, *7*(*3*), 300–312.

Meyers, J. E., & Meyers, K. R. (1995). *Rey complex figure test and recognition trial (RCFT)*. Odessa, FL: Psychological Assessment Resources.

Miller, L. A., & Lovler, R. L. (2018). *Foundations of psychological testing: A practical approach*. Thousand Oaks, CA: Sage Publications.

Morris, J. C. (1993). The clinical dementia rating (CDR): Current version and scoring rules. *Neurology*, *43*(*11*), 2412.

Morris, J. C. (1997). Clinical dementia rating: A reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics*, *9*(*S1*), 173–176.

Munro, C. A., Winicki, J. M., Schretlen, D. J., Gower, E. W., Turano, K. A., Muñoz, B., et al. (2012). Sex differences in cognition in healthy elderly individuals. *Aging, Neuropsychology, and Cognition*, *19*(*6*), 759–768. doi: 10.1080/13825585.2012.690366.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(*4*), 695–699.

O'Bryant, S. E., Lacritz, L. H., Hall, J., Waring, S. C., Chan, W., Khodr, Z. G., et al. (2010). Validation of the new interpretive guidelines for the clinical dementia rating scale sum of boxes score in the national Alzheimer's coordinating center database. *Archives of Neurology*, *67*(*6*), 746–749.

O'Bryant, S. E., Waring, S. C., Cullum, C. M., Hall, J., Lacritz, L., Massman, P. J., et al. (2008). Staging dementia using clinical dementia rating scale sum of boxes scores: A Texas Alzheimer's research consortium study. *Archives of Neurology*, *65*(*8*), 1091–1095.

Oltra-Cucarella, J., Sánchez-SanSegundo, M., Rubio-Aparicio, M., Arango-Lasprilla, J. C., & Ferrer-Cascales, R. (2019). The association between the number of neuropsychological measures and the base rate of low scores. *Assessment*, e-publication ahead of print. doi: 10.1177/1073191119864646.

Papp, K. V., Buckley, R., Mormino, E., Maruff, P., Villemagne, V. L., Masters, C. L., et al. (2019). Clinical meaningfulness of subtle cognitive decline on longitudinal testing in preclinical AD. *Alzheimer's & Dementia*, *16*(*3*), 552–560.

Partington, J. E., & Leiter, R. G. (1949). Partington pathways test. *Psychological Service Center Journal*, *1*, 11–20.

Porto, M.-F., Russo, M.-J., & Allegri, R. (2018). Neuropsychological battery uniform data set (UDS) for the evaluation of Alzheimer's disease and mild cognitive impairment: A systematic review. *Revista Ecuatoriana de Neurologia*, *27*(*2*), 55–62.

Possin, K. L., Laluz, V. R., Alcantar, O. Z., Miller, B. L., & Kramer, J. H. (2011). Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer's disease and behavioral variant frontotemporal dementia. *Neuropsychologia*, *49*(*1*), 43–48.

Rapport, L. J., Brines, D. B., Theisen, M. E., & Axelrod, B. N. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist*, *11*(*4*), 375–380.

Royall, D. R., & Palmer, R. F. (2014). "Executive functions" cannot be distinguished from general intelligence: Two variations on a single theme within a symphony of latent variance. *Frontiers in Behavioral Neuroscience*, *8*, 369.

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, *16*(*5*), 754–760.

Samtani, M. N., Raghavan, N., Novak, G., Nandy, P., & Narayan, V. A. (2014). Disease progression model for clinical dementia rating–sum of boxes in mild cognitive impairment and Alzheimer's subjects from the Alzheimer's disease Neuroimaging initiative. *Neuropsychiatric Disease and Treatment*, *10*, 929.

Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, *8*(*3*), 448–460.

Thomas, K. R., Bangen, K. J., Weigand, A. J., Edmonds, E. C., Wong, C. G., Cooper, S., et al. (2019). Objective subtle cognitive difficulties predict future amyloid accumulation and neurodegeneration. *Neurology*, *94*(*4*), e397–e406.

Thomas, K. R., Edmonds, E. C., Eppig, J., Salmon, D. P., Bondi, M. W., & Alzheimer's Dis Neuroimaging I (2018). Using neuropsychological process scores to identify subtle cognitive decline and predict progression to mild cognitive impairment. *Journal of Alzheimers Disease*, *64*(*1*), 195–204. doi: 10.3233/jad-180229.

Weintraub, S., Besser, L., Dodge, H. H., Teylan, M., Ferris, S., Goldstein, F. C., et al. (2018). Version 3 of the Alzheimer disease centers' neuropsychological test battery in the uniform data set (UDS). *Alzheimer Disease and Associated Disorders*, *32*(*1*), 10.

White, T., & Stern, R. A. (2003). *Neuropsychological assessment battery psychometric and technical manual*. Lutz, FL: Psychological Assessment Resources.

Woods, S. P., Childers, M., Ellis, R. J., Guaman, S., Grant, I., Heaton, R. K., et al. (2006). A battery approach for measuring neuropsychological change. *Archives of Clinical Neuropsychology*, *21*(*1*), 83–89.