Journal of the International Neuropsychological Society (2015), **21**, 558–567.

doi:10.1017/S1355617715000582

Reliable Change on Neuropsychological Tests in the Uniform Data Set

Brandon E. Gavett,[1] Lee Ashendorf,[2] AND Ashita S. Gurnani[1]

[1]University of Colorado, Colorado Springs, Department of Psychology, Colorado Springs, Colorado
[2]Boston University School of Medicine, Department of Psychiatry, Boston, Massachusetts

(RECEIVED October 7, 2014; FINAL REVISION June 16, 2015; ACCEPTED June 26, 2015; FIRST PUBLISHED ONLINE August 3, 2015

Abstract

Longitudinal normative data obtained from a robust elderly sample (i.e., believed to be free from neurodegenerative disease) are sparse. The purpose of the present study was to develop reliable change indices (RCIs) that can assist with interpretation of test score changes relative to a healthy sample of older adults (ages 50+). Participants were 4217 individuals who completed at least three annual evaluations at one of 34 past and present Alzheimer's Disease Centers throughout the United States. All participants were diagnosed as cognitively normal at every study visit, which ranged from three to nine approximately annual evaluations. One-year RCIs were calculated for 11 neuropsychological variables in the Uniform Data Set by regressing follow-up test scores onto baseline test scores, age, education, visit number, post-baseline assessment interval, race, and sex in a linear mixed effects regression framework. In addition, the cumulative frequency distributions of raw score changes were examined to describe the base rates of test score changes. Baseline test score, age, education, and race were robust predictors of follow-up test scores across most tests. The effects of maturation (aging) were more pronounced on tests related to attention and executive functioning, whereas practice effects were more pronounced on tests of episodic and semantic memory. Interpretation of longitudinal changes on 11 cognitive test variables can be facilitated through the use of reliable change intervals and base rates of score changes in this robust sample of older adults. A Web-based calculator is provided to assist neuropsychologists with interpretation of longitudinal change. (*JINS*, 2015, *21*, 558–567)

Keywords: Reliability of results, Longitudinal studies, Cognition, Cognitive symptoms, Aging, Dementia

INTRODUCTION

Neuropsychologists are often tasked with re-evaluating individuals to help determine whether cognitive functioning has changed over a given time interval. Most neuropsychological test instruments are interpreted using normative data collected from a putatively healthy sample to understand the expected mean and variance in test scores produced by nondiseased persons (Mitrushina, Boone, Razani, & D'Elia, 2005). These normative data are typically used to interpret an individual person's test scores in the context of his or her peers, with corrections for demographic factors such as age, education, sex, and race (Heaton, Miller, Taylor, & Grant, 2004). When applying norms for the purpose of understanding change in older adults, there are two critical issues that could undermine interpretation.

First, it is difficult to determine whether the normative data are robust to latent causes of cognitive difficulties, especially in older age groups. An older person who is part of a normative sample may be in the very early stages of a neurodegenerative disease, such as Alzheimer's disease, but may not be manifesting clinically obvious cognitive difficulties at the time the normative data were collected. Recent efforts have been made to include participants believed to be disease-free after several years of follow-up ("robust norms;" Holtzer et al., 2008; Pedraza et al., 2010), as a means of ensuring that the normative sample is representative of cognitively healthy individuals.

Second, norms are generally cross-sectional in nature, not longitudinal, yet are interpreted to reflect magnitude of change when used for repeated assessments of patients or research participants. This ignores properties of the test such as reliability and practice effects, and it also discounts statistical effects such as regression to the mean (McCaffrey, Duff, & Westervelt, 2000). Various statistical methods have been proposed to account for these potential confounds, ranging from simple standard deviation difference methods

Correspondence and reprint requests to: Brandon E. Gavett, UCCS Department of Psychology, 1420 Austin Bluffs Parkway, Colorado Springs, CO 80918. E-mail: bgavett@uccs.edu

(see Frerichs & Tuokko, 2005), to reliable change models of varying complexity (see Hinton-Bayre, 2010), to standardized regression-based (SRB) methods (e.g., Attix et al., 2009). See Duff (2012) and Heilbronner et al. (2010) for a more detailed discussion of these and other issues related to serial assessment in neuropsychology. Robust longitudinal norms contextualize the change in an individual's test scores relative to a sample that is believed to have been free from neurodegenerative disease during the test–retest interval. Change in test scores that is more extreme than that observed in robust normative samples may reflect a change in cognition that is beyond the limits of normal aging (Bläsi et al., 2009). Robust norms have been shown to improve diagnostic accuracy in the longitudinal assessment of older adults (De Santi et al., 2008; Holtzer et al., 2008; Pedraza et al., 2010).

In this study, we propose to address the two weaknesses discussed above by quantifying expected changes in cognitive abilities over time through the use of linear mixed effects regression models to calculate reliable change intervals (RCIs). Linear mixed effects models are an extension of SRB models, which address longitudinal data by allowing for individual variability in baseline test scores (intercepts) and rate of change over time (slopes; Pinhiero & Bates, 2000). These models can be used to predict an examinee's follow-up test score based on variables such as the examinee's baseline test score and several demographic variables. The observed follow-up test score is compared to the predicted follow-up test score, and if the difference is large enough, the change may be interpreted as reliable. The magnitude of reliable change is scaled relative to the standard error observed in the linear mixed effects model and the degree of confidence desired in the prediction interval (often 90%). For instance, if the standard error is 2.0 and the desired degree of confidence for the interval is 90%, then the confidence interval would have a range of 2.0 times the standard normal distribution quantile associated with a two-tailed alpha level of .05 (i.e., 1.645). (For small sample sizes, this standard normal quantile can be replaced with the appropriate *t* distribution quantile for a given degrees of freedom.) In this example, $2 \times 1.645 = 3.29$, indicating that the 90% confidence interval would have range of 3.29 units in both the positive and negative directions. Differences between observed and predicted follow-up scores that are more extreme than ± 3.29 are thus suggestive of reliable change. By applying RCIs to neuropsychological measurements, one can identify whether a change in a given score is clinically interpretable. We seek to produce robust longitudinal change indices that can be used *in vivo* to determine whether individuals are changing at a rate that is consistent with normal aging, whether an individual's rate of change is more rapid than expected, or whether a treatment has a beneficial effect on cognition.

We will identify individuals from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS; Beekly et al., 2007; Morris et al., 2006) who have been confirmed through at least three (and up to nine) longitudinal clinical assessments to be cognitively healthy. We will then retrospectively examine the first (baseline) and second (follow-up) visits to quantify the degree of change observed across time in this putatively healthy sample. While the UDS neuropsychological battery is well established (Weintraub et al., 2009), the psychometric properties are still under evaluation and no RCIs have yet been presented, limiting the effectiveness and potentially the accuracy of longitudinal evaluations using this selection of tests. As the UDS neuropsychological battery is possibly the most widely used research battery for the cognitive assessment of dementia in the United States, it is important to identify the longitudinal psychometric characteristics of this battery, for both research and clinical purposes. The objective of this study is to present RCIs based on linear mixed effects models for each of the available UDS neuropsychological variables. As a result, readers will have access to robust longitudinal data that can be used to interpret cognitive changes in older adults.

METHOD

Participants

This study was determined to be exempt from human subjects review by the University of Colorado, Colorado Springs Institutional Review Board. Data used in the present study were obtained from the NACC's publicly available database. Created by the National Institute on Aging, the NACC compiles a wide variety of data, including neuropsychological test scores from 34 past and present Alzheimer's Disease Centers (ADCs) using the UDS battery. We included participants who had completed at least three visits, including one baseline visit, between September 2005 and March 2014. A total of 4598 individuals in the database were diagnosed as cognitively normal at all visits. We also excluded 92 participants who were less than 50 years old at their baseline visit and 302 participants who did not speak English as their primary language or who were not assessed in English. In total, we excluded 381 participants (13 participants met more than one of the exclusion criteria), leaving a sample of 4217 for inclusion in the study. These participants underwent at least three—and up to nine—approximately annual evaluations at an ADC and were diagnosed as cognitively normal at all evaluations. Because very few participants completed more than seven visits, we did not analyze data from the eighth or ninth visits. See Table 1 for details regarding participant demographic variables.

Measures

The neuropsychological measures available for these analyses included the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), Wechsler Adult Intelligence Scale-Revised (WAIS-R) Digit Span Forward and Backward conditions (Wechsler, 1981), WAIS-R Digit Symbol (Wechsler, 1981), Trail Making Test (TMT) parts A and B (Reitan & Wolfson, 1993), Story A from Wechsler Memory Scale-Revised (WMS-R) Logical Memory (Wechsler, 1987), two semantic fluency tasks (animals and vegetables; Weintraub et al., 2009), and the 30 odd-item short

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

form of the Boston Naming Test (BNT) (Jefferson et al., 2007). These tests are common to most dementia clinicians and researchers, and will not be described here. See the study by Weintraub and colleagues (2009) for more information.

Data Analysis

The test data, which were available for as few as three to as many as seven approximately annual visits, were used in a linear mixed effects model for each test, with visit number nested within participants. For each test, we modeled linear, quadratic, and logarithmic trends and found that a linear trend provided the best balance between model fit and parsimony (data not shown). Reliable change intervals were derived for the second visit only. All analyses were performed in R version 3.1.2 (R Core Team, 2015). The *lme4* package (version 1.1-8) was used for longitudinal modeling (Bates, Maechler, Bolker, & Walker, 2015).

Eleven linear mixed effects regression models, one for each test, were specified to include both fixed and random (intercept and slope) effects. The follow-up test scores from visits two to seven were regressed onto the following fixed effects: baseline test score, age at baseline (years), education (years), visit number, assessment interval (years post-baseline), race (Caucasian or non-Caucasian), and sex (male or female). All predictor variables were entered simultaneously; although stepwise regression procedures have been used previously for RCI studies in the neuropsychology literature, these methods were not used here. Being data driven, rather than theory driven, models identified using stepwise methods have the potential to capitalize on chance and may not generalize beyond the sample data; many other limitations have also been identified (e.g., Whittingham, Stephens, Bradbury, & Freckleton, 2006). Dummy coding was used for race and sex, with Caucasians and males as the reference categories for their respective groups. For each model, fixed effects parameter estimates and their 95% confidence intervals were obtained using restricted maximum likelihood estimation. The standard deviation of the random intercepts and slopes were also obtained. Predicted follow-up scores were based on the fixed effects parameter estimates only. To account for the variability introduced by the uncertainty in both the fixed and random effects, 90% reliable change intervals were based on the residual standard error as well as the variability in the predictions. The variability in the predictions was estimated using parametric bootstrapping ($B = 1000$) of the predicted test scores across all visits. This bootstrapping procedure was based on simulated values for the random effects to account for these sources of variability and results in unique prediction intervals for each participant.

In addition to calculating RCI, we also examined the frequency with which raw scores changed from baseline to follow-up. To establish base rates for longitudinal change in this sample, we derived cumulative percentages for raw score changes of each observed magnitude. It is not uncommon for "statistically significant" score differences to occur frequently in healthy samples (Matarazzo & Herman, 1984). Therefore, these base rate data can serve to augment the RCI values to not only determine the statistical significance of the observed change from baseline to follow-up, but to determine the relative frequency of change of a given magnitude.

RESULTS

Participant demographics are presented in Table 1. Descriptive statistics for the 11 neuropsychological tests at baseline and 1-year follow-up are presented in Table 2. The fixed effects parameter estimates and their 95% confidence intervals are presented in Table 3, along with the standard deviations of the random effects. For each test, the random slope accounted for very little variability, with *SD*s ranging from 0.09 (MMSE) to 3.99 (TMT-B); in contrast, the *SD*s of the random intercept terms were more sizeable, ranging from 0.66 (MMSE) to 20.99 (TMT-B). These results suggest that, although individuals varied in their baseline test scores, there is little heterogeneity in individual trajectories of change over time on any of the tests. These patterns of change are depicted graphically in Figure 1. As seen in this figure, the margin of error in the average reliable change intervals increases, sometimes asymmetrically, across visits for most tests. A closer examination of the fixed effects parameter estimates and their 95% confidence intervals in Table 3 reveals that, for most tests, baseline test score, age, education, and race were the most reliable predictors of follow-up test score. Higher baseline test scores, younger age, more years of education, and Caucasian race were associated with better performance on all follow-up test scores. Female sex was associated with higher follow-up scores on the MMSE, Digit Symbol Coding, vegetable fluency, and Logical Memory I and II, whereas male sex was predictive of higher follow-up scores on the BNT. A longer post-baseline interval was predictive of worse follow-up scores on all tests except the MMSE and Forward Digit Span. More frequent exposure to tests (i.e., a larger number of previous visits) yielded better scores on Backward Digit Span, Digit Symbol Coding, the BNT, and the two Logical Memory subtests. Neither visit number nor

Table 1. Participant demographics

Variable	*N*	*M*	*SD*	Range
Visits	4217	5.12	1.68	3–9
Age at baseline	4217	72.61	8.77	50.1–100.3
Education (years)	4202	15.80	2.79	3–25
Sex (Female)	2857 (67.7%)	—	—	—
Caucasian race	3440 (81.6%)	—	—	—
Hispanic ethnicity	88 (2.1%)	—	—	—
T1 to T2 Interval (months)	4217	14.62	5.20	4.8–63.6
T2 to T3 Interval (months)	4217	13.92	4.73	3.6–58.8
T3 to T4 Interval (months)	3256	13.51	3.79	2.4–55.2
T4 to T5 Interval (months)	2414	13.25	3.50	4.8–51.6
T5 to T6 Interval (months)	1771	12.86	2.98	1.2–33.6
T6 to T7 Interval (months)	1056	12.59	2.24	7.2–26.4

Note. *N* = sample size; *M* = mean; *SD* = standard deviation.

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

Table 2. Descriptive statistics for each UDS test at baseline and 1-year follow-up

	Baseline visit						Follow-up visit						Test–retest reliability	
Test	*M*	*SD*	S	K	Min	Max	*M*	*SD*	S	K	Min	Max	*r*	95% CI
MMSE	29.06	1.27	–2.09	7.16	17.00	30.00	29.05	1.36	–2.28	8.03	18.00	30.00	0.49	[.47, .52]
DS-F	8.73	2.02	–0.21	–0.61	1.00	12.00	8.73	2.00	–0.20	–0.50	1.00	12.00	0.66	[.64, .67]
DS-B	7.02	2.24	0.23	–0.45	0.00	12.00	7.04	2.24	0.25	–0.49	1.00	12.00	0.67	[.66, .69]
Digit Symbol	48.61	12.00	–0.05	0.33	3.00	91.00	48.97	12.33	–0.03	0.23	3.00	93.00	0.86	[.85, .87]
TMT-A	33.64	14.78	2.51	11.23	12.00	150.00	32.87	14.47	2.67	13.67	11.00	150.00	0.64	[.62, .66]
TMT-B	86.13	47.13	2.29	6.62	25.00	300.00	85.03	47.21	2.40	7.22	24.00	300.00	0.74	[.73, .76]
Animals	20.64	5.65	0.29	0.09	1.00	52.00	20.60	5.62	0.33	0.44	0.00	50.00	0.67	[.65, .69]
Vegetables	15.20	4.22	0.29	0.39	1.00	36.00	15.16	4.27	0.30	0.13	2.00	31.00	0.61	[.59, .63]
BNT	27.36	3.18	–2.67	10.21	0.00	30.00	27.61	3.12	–2.94	12.87	0.00	30.00	0.78	[.77, .79]
LM-I	13.98	3.77	–0.17	–0.15	0.00	24.00	14.49	3.76	–0.21	–0.13	1.00	25.00	0.62	[.60, .64]
LM-D	12.78	4.05	–0.12	–0.19	0.00	24.00	13.43	4.07	–0.22	–0.12	0.00	25.00	0.65	[.63, .67]

Note. *M* = mean; *SD* = standard deviation; S = skewness; K = kurtosis; MMSE = Mini-Mental State Examination; DS-F = Digit Span Forward; DS-B = Digit Span Backward; TMT = Trail Making Test; BNT = Boston Naming Test; LM-I = Logical Memory Immediate Recall; LM-D = Logical Memory Delayed Recall.

post-baseline interval were predictive of follow-up scores on the MMSE and Forward Digit Span.

As there is concern for potential heteroscedasticity among the regression residuals, a plot of the residuals *versus* fitted values is provided in Figure 2. The models for MMSE, TMT-A, TMT-B, and BNT should be interpreted with caution due to non-normal score distributions caused by floor and ceiling effects. Floor effects (for the TMT) and ceiling effects (for MMSE and BNT) in the data may bias the interpretation of change scores in examinees who are close to floor or ceiling on these tests at baseline.

Table 4 contains data relevant to the reliable change indices from baseline to the first annual follow-up visit. The methods used in this study produce a unique RCI for each participant. To summarize the margin of error needed for reliable change, the data shown in Table 4 were derived from the average participant in our sample [i.e., with mean values of all continuous predictor variables and modal values for sex (i.e., female) and race (i.e., Caucasian)]. The column labeled "SEE" reflects the residual standard error, as reported in Table 3. The column labeled "90% PI MOE" represents the bootstrapped margin of error for predicting follow-up test scores, conditioned on all random effects. The column labeled "90% RCI MOE" represents the margin of error for the 90% reliable change intervals. If the difference between observed and predicted follow-up scores falls outside of this interval, the change may be interpreted as reliable with 90% confidence. The test scores associated with several relevant base rates of score changes on these 11 tests are presented in Table 5.

Readers wishing to obtain reliable change intervals for other combinations of predictor variables are referred to the Web-based calculator created to supplement this manuscript. It should be noted, however, that predictions for out-of-sample data cannot be conditioned on the random effects, which may underestimate the magnitude of the reliable change intervals. This calculator can be accessed at https://begavett.shinyapps.io/UDS_RCI.

DISCUSSION

As the aging population continues to grow worldwide, the number of individuals who suffer from neurodegenerative diseases also continues to grow (Sosa-Ortiz, Acosta-Castillo, & Prince, 2012). Clinical diagnosis of neurodegenerative disease requires a change from a baseline level of functioning (McKhann et al., 2011), which supports the need for serial assessment. Despite the clear importance of serial assessment in the tracking of longitudinal cognitive decline, relatively little attention has been paid to issues of interpreting change scores. Without an understanding of factors such as normal aging, practice effects, regression to the mean, and measurement error, it may be easy to misinterpret score differences between baseline and follow-up. Because there are very limited normative data available for serial assessment data and change scores, interpretation of change is often subjective.

The current study adds to the reliable change literature in two important ways. First, we have used linear mixed effects regression to model change in cognitive test scores over at least three and as many as seven approximately annual visits. The results of these analyses reveal that there is little heterogeneity in the individual trajectories of change over time in a large sample believed to be free from cognitive impairment. Second, these results also help to tease apart the relative contributions of maturation (i.e., normal aging) and practice effects that can affect follow-up test scores. Of the 11 test scores examined here, practice effects were most evident for Backward Digit Span, Digit Symbol Coding, the BNT, and the two Logical Memory subtests. Based on the parameter estimates for these tests, a one-point test score increase appears after approximately 2 visits for Logical Memory Immediate and Delayed, 3 visits for Digit Symbol Coding, 9 visits for the BNT, and 17 visits for Backward Digit Span, when holding all other predictor variables constant. For many tests, these practice effects are outweighed by the length of the post-baseline assessment interval, which was inversely

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

Table 3. Linear mixed effects regression parameter estimates for predicting follow-up test scores across seven annual visits

	Fixed effects parameter estimates [95% CI]								*SD* of random effects		
Test	Intercept	Baseline score	Age	Education	Visit #	Post-baseline interval (years)	Race	Sex	Intercept	Slope	Residual
MMSE	17.91 [17.19, 18.63]	0.41 [0.39, 0.43]	−0.02 [−0.03, −0.02]	0.07 [0.06, 0.08]	−0.01 [−0.06, 0.04]	−0.01 [−0.05, 0.04]	−0.71 [−0.78, −0.64]	0.21 [0.15, 0.27]	0.66	0.09	0.93
DS-F	3.72 [3.28, 4.17]	0.60 [0.59, 0.62]	−0.01 [−0.02, −0.01]	0.06 [0.04, 0.07]	−0.02 [−0.08, 0.05]	0.01 [−0.04, 0.07]	−0.28 [−0.38, −0.19]	−0.06 [−0.13, 0.02]	1.06	0.16	1.13
DS-B	3.54 [3.06, 4.02]	0.61 [0.60, 0.63]	−0.02 [−0.03, −0.02]	0.05 [0.04, 0.07]	0.10 [0.03, 0.17]	−0.04 [−0.10, 0.02]	−0.55 [−0.66, −0.45]	0.01 [−0.07, 0.10]	1.18	0.16	1.24
DSC	20.86 [18.56, 23.16]	0.82 [0.80, 0.84]	−0.21 [−0.23, −0.19]	0.16 [0.10, 0.23]	0.76 [0.48, 1.04]	−0.52 [−0.76, −0.28]	−1.68 [−2.13, −1.23]	0.95 [0.59, 1.32]	4.66	0.77	4.19
TMT-A	−7.19 [−10.31, −4.07]	0.52 [0.50, 0.54]	0.37 [0.34, 0.40]	−0.31 [−0.41, −0.21]	−0.41 [−0.90, 0.07]	0.34 [−0.07, 0.75]	4.39 [3.65, 5.13]	0.01 [−0.57, 0.60]	6.56	0.97	7.81
TMT-B	−21.97 [−30.99, −12.95]	0.64 [0.62, 0.66]	0.94 [0.84, 1.04]	−1.30 [−1.60, −1.01]	−1.03 [−2.49, 0.42]	2.94 [1.71, 4.18]	14.68 [12.47, 16.88]	−0.46 [−2.14, 1.22]	20.99	3.99	24.06
Animals	11.65 [10.40, 12.90]	0.56 [0.54, 0.58]	−0.09 [−0.10, −0.08]	0.23 [0.20, 0.27]	0.23 [0.05, 0.41]	−0.15 [−0.30, 0.01]	−1.39 [−1.66, −1.12]	0.16 [−0.05, 0.38]	2.80	0.42	3.08
Vegetables	10.14 [9.15, 11.13]	0.51 [0.49, 0.53]	−0.08 [−0.09, −0.07]	0.15 [0.12, 0.18]	0.13 [−0.02, 0.28]	−0.10 [−0.23, 0.03]	−0.94 [−1.15, −0.73]	1.40 [1.22, 1.59]	2.17	0.37	2.55
BNT	10.03 [9.27, 10.79]	0.69 [0.67, 0.71]	−0.03 [−0.04, −0.03]	0.06 [0.05, 0.08]	0.19 [0.11, 0.27]	−0.03 [−0.10, 0.04]	−0.70 [−0.84, −0.56]	−0.17 [−0.27, −0.07]	1.52	0.25	1.30
LM-I	7.19 [6.34, 8.04]	0.53 [0.51, 0.55]	−0.05 [−0.06, −0.04]	0.14 [0.11, 0.16]	0.84 [0.71, 0.97]	−0.27 [−0.38, −0.16]	−1.40 [−1.59, −1.21]	0.68 [0.53, 0.84]	2.17	0.35	2.13
LM-D	6.81 [5.91, 7.71]	0.56 [0.54, 0.58]	−0.06 [−0.07, −0.05]	0.14 [0.11, 0.17]	1.10 [0.96, 1.23]	−0.27 [−0.39, −0.16]	−1.64 [−1.85, −1.44]	0.53 [0.36, 0.70]	2.24	0.38	2.17

Note. CI = Confidence interval; *SD* = Standard Deviation; MMSE = Mini-Mental State Examination; DS-F = Digit Span Forward; DS-B = Digit Span Backward; DSC = Digit Symbol Coding; TMT = Trail Making Test; BNT = Boston Naming Test; LM-I = Logical Memory Immediate Recall; LM-D = Logical Memory Delayed Recall.

associated with performance on Backward Digit Span, Digit Symbol Coding, TMT-A and B, both semantic fluency tasks, the BNT, and both Logical Memory subtests. For Backward Digit Span, Digit Symbol Coding, TMT-A and B, and semantic fluency, the post-baseline assessment interval had a more pronounced effect than the influence of practice. On the other hand, practice effects outweighed maturation effects on the BNT and both Logical Memory subtests. Therefore, literature on practice effects may be augmented by consideration of test–retest intervals (e.g., Duff, Callister, Dennett, & Tometich, 2012).

These linear mixed effects models were used to calculate a standard error for predicted test scores at an examinee's second visit. These standard errors are used, along with the variability in predicted test scores, to generate 90% reliable change intervals, which provide a range of difference scores that fall within the test's margin of error while accounting for several important covariates and sources of variability. The results provide empirical data on change scores from baseline to approximately 1-year follow-up in a robust sample of participants who underwent at least three approximately annual evaluations and were never diagnosed with any form of cognitive impairment at any visit. Using regression methods that account for maturation effects (i.e., aging), practice effects, regression to the mean, baseline test scores, and demographic variables, we present data for eleven different UDS neuropsychological test variables that can be used to calculate a predicted follow-up test score and 90% reliable change intervals for the difference between observed and predicted follow-up scores. Follow-up test score changes that fall outside of these intervals can be interpreted as reflecting "true" change with a magnitude that is larger than would be expected based on the measurement error of the test. To augment these reliable change intervals, we also present data on the frequency with which score changes were observed in this robust sample. Because statistically significant changes in test scores may often be very frequent in a clinical sample, interpreting RCIs along with base rate data can assist with the interpretation of score changes in the context of how commonly or rarely such a change score is expected to occur in a normative sample.

By way of an example, consider a 73-year-old, college-educated Caucasian man evaluated using these UDS measures, with scores and percentiles (calculated using Shirk et al., 2011) presented in the first two columns of Table 6. If we were to determine "impairment" by using a global cutoff of $Z = -1.5$ (7^{th} percentile), we would find no scores below that cutoff and therefore there are no impaired cognitive domains at this initial visit. Thirteen months later, he is seen for his first follow-up, reports no functional problems, and his neuropsychological test scores are provided in the second two columns of that table. Using the same standard for "impairment," we would say that he is now impaired on Digit Symbol Coding and TMT-B and exhibits difficulty with complex processing speed. However, using the RCIs developed above and as obtained from the Web-based calculator, we can see that he exhibited decline in excess of the 90%

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

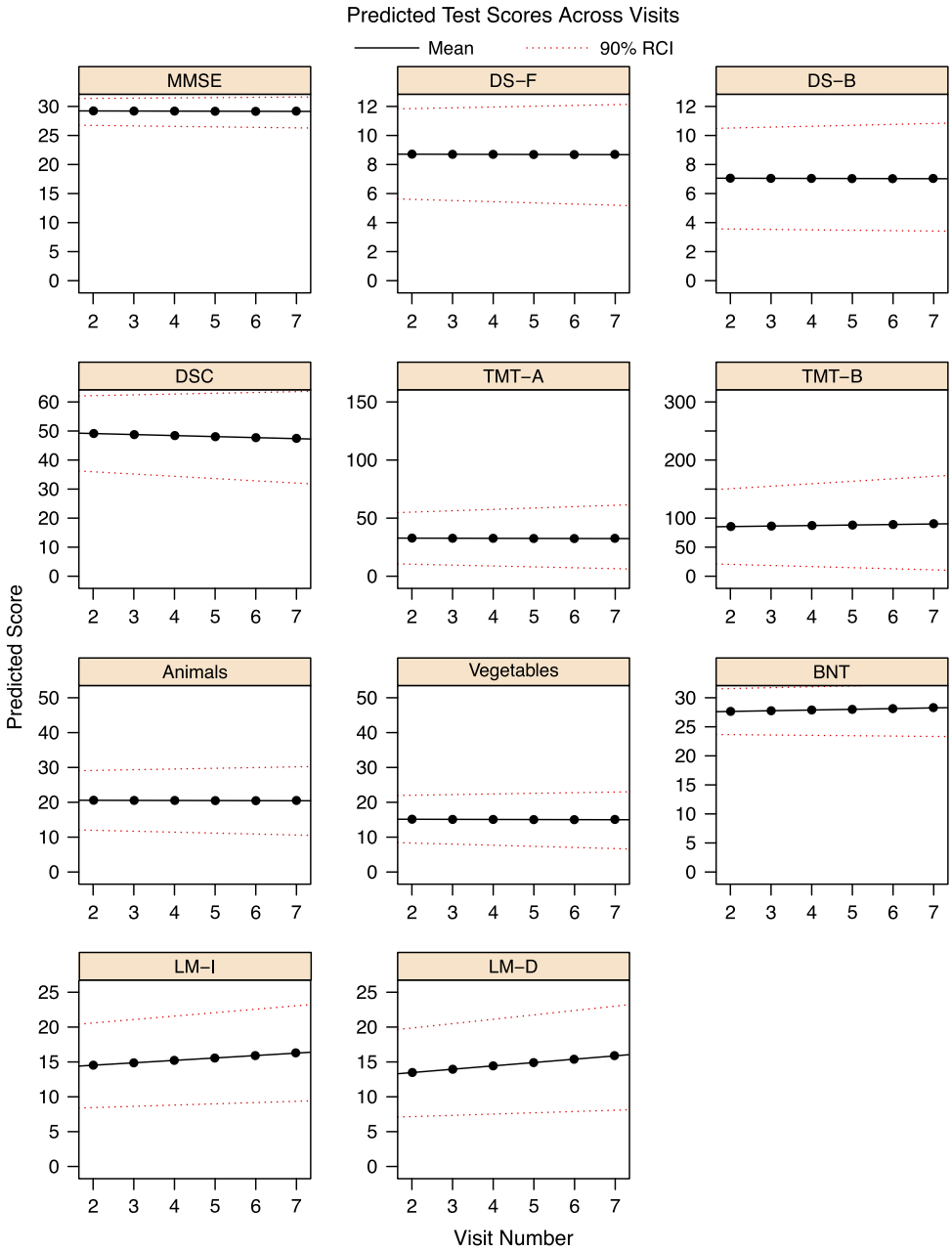

Fig. 1. Predicted test scores (black circles) and 90% reliable change intervals (dotted red lines) for each test across visits 2 to 7 based on linear mixed effects regression.

interval of change (change *Z*-score $> \pm 1.645$) on MMSE, animal naming, vegetable fluency, and the BNT. Even though he is not "impaired" in the language domain using the fixed *Z*-score criterion, he displayed decline on three language-domain tasks relative to the baseline exam, suggesting that this might be a domain of clinical interest. In contrast, while Digit Symbol Coding and TMT-B both technically declined into the impaired range, neither test showed reliable change across visits and therefore, despite the newly developed "impairment" on these tests, this cannot be interpreted as a decline relative to the visit 1 baseline.

All demographic variables were found to contribute to the prediction of follow-up scores, with some (e.g., age, education) more robust than others (e.g., sex). It should be noted that these results were obtained from a sample of older adults who were diagnosed as cognitively healthy at their baseline visit. Therefore, the results presented in this study, especially the data used to predict follow-up test scores (Table 3) cannot be generalized beyond this population. It would be a misuse of the data to attempt to predict follow-up test scores for individuals with cognitive impairment at baseline. Similarly, the results will not generalize to individuals whose baseline test scores are not included in the test score intervals

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

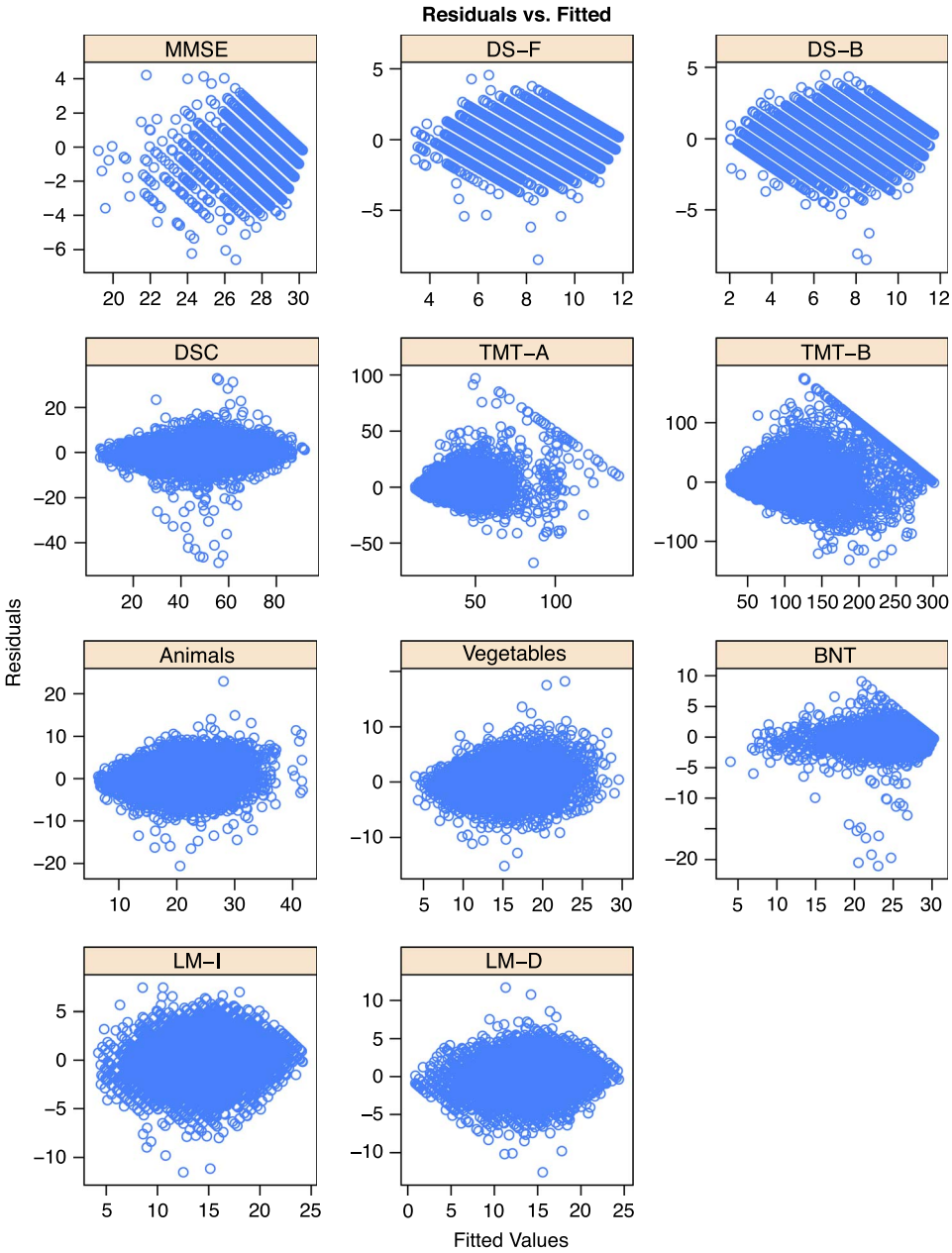

Fig. 2. Residuals *versus* fitted plots at visit 2 for each test based on linear mixed effects regression.

presented in Table 2, or to people whose demographic variables or test–retest intervals were not observed in the current study.

This study is limited in several ways. First, the data in the current sample were obtained from the NACC, which compiles data from 34 past and present ADCs across the United States. Each ADC may differ somewhat in its recruitment methods, especially for cognitively healthy individuals. The sample used in this study was not recruited for the purposes of producing normative data (e.g., random sampling was not used), and valid concerns may be raised about the external validity of these findings. The sample was also highly educated ($M = 15.80$; $SD = 2.79$) and was under-representative of racial and ethnic minorities. In contrast, the sample is very large, geographically diverse, and continued follow-up beyond the two visits used in this study gives confidence that the participants were not in the early stages of a neurodegenerative disease at the time the data were collected. Several of the neuropsychological test variables in the UDS have non-normal distributions. As discussed above, truncated distributions may be associated with heteroscedasticity (Figure 2), which could contribute to an underestimate of the residual variance for tests with floor or ceiling effects (i.e., MMSE, TMT, BNT).

Another limitation of the results is the finding that most of the test variables included in the current study possessed

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

Table 4. Reliable change intervals from baseline to the first annual follow-up visit for the average participant in the study sample

Test	SEE	90% PI MOE	90% RCI MOE
MMSE	0.93	[−0.67, 1.07]	[−2.20, 2.60]
DS-F	1.13	[−1.77, 0.84]	[−3.62, 2.69]
DS-B	1.24	[−0.83, 2.20]	[−2.87, 4.25]
Digit Symbol	4.19	[−12.27, 0.55]	[−19.16, 7.44]
TMT-A	7.81	[−6.15, 13.05]	[−19.00, 25.90]
TMT-B	24.06	[−15.26, 38.10]	[−54.83, 77.67]
Animals	3.08	[−4.15, 3.55]	[−9.23, 8.63]
Vegetables	2.55	[−5.86, −0.45]	[−10.06, 3.75]
BNT	1.30	[−2.01, 1.69]	[−4.14, 3.82]
LM-I	2.13	[−2.43, 2.90]	[−5.93, 6.40]
LM-D	2.17	[−3.92, 1.78]	[−7.49, 5.35]

Note. SEE = standard error of the estimate; PI MOE = Prediction Interval Margin of Error; RCI MOE = Reliable Change Interval Margin of Error; MMSE = Mini-Mental State Examination; DS-F = Digit Span Forward; DS-B = Digit Span Backward; TMT = Trail Making Test; BNT = Boston Naming Test; LM-I = Logical Memory Immediate Recall; LM-D = Logical Memory Delayed Recall.

test–retest reliabilities below.70 (Table 2). These findings are roughly consistent with 1-year test–retest reliability estimates derived from meta-analysis (Calamia, Markon, & Tranel, 2013). The change in mean scores from baseline to follow-up is likely to be reflective of the magnitude of history and maturational influences acting across the two time points. The strength of the correlation between test scores at two successive time points may be indicative of individual differences in variability of change (Salthouse & Tucker-Drob, 2008). The low test–retest correlations could be attributed to random error, real change in the construct validity of the test between the two time points, or measurement error. Although maturational influences may affect within-person change in test scores, we also show that practice effects contribute to change in performance on most tests (Salthouse & Tucker-Drob, 2008). As might be expected, tests involving attention, processing speed, mental efficiency, and working memory were more susceptible to maturational influences (i.e., longer test–retest intervals), whereas tests involving episodic and semantic memory were more susceptible to practice effects.

Table 5. UDS test score changes from baseline to the first annual follow-up visit corresponding to various base rates

	Base rates (cumulative percentages)						
Test	1%	5%	16%	50%	84%	95%	99%
MMSE	−4	−2	−1	0	1	2	3
DS-F	−4	−3	−2	0	2	3	4
DS-B	−4	−3	−2	0	2	3	4
DSC	−14	−9	−5	0	5	9	16
TMT-A	28	15	7	−1	−9	−17	−34
TMT-B	100	46	20	−1	−22	−50	−107
LM-I	−7	−5	−3	1	4	6	8
LM-D	−7	−5	−3	1	4	6	8
Animals	−11	−8	−4	0	4	8	11
Vegetables	−9	−6	−4	0	4	6	9
BNT	−4	−2	−1	0	2	3	6

Note. MMSE = Mini-Mental State Examination; DS-F = Digit Span Forward; DS-B = Digit Span Backward; DSC = Digit Symbol Coding; TMT = Trail Making Test; LM-I = Logical Memory Immediate Recall; LM-D = Logical Memory Delayed Recall; BNT = Boston Naming Test.

Although a minority of the tests in the UDS battery are current and in common use in clinical settings (i.e., TMT, animal fluency, BNT), these results may still be valuable to both clinicians and researchers who perform cognitive evaluations of older adults. While newer editions of these tests have been published in recent years (e.g., the WAIS and WMS have twice been updated), it is unclear whether these updates have led to substantial improvements in the longitudinal measurement properties of these tests for the assessment of elderly individuals. The results of the current study can be valuable in that there is a paucity of longitudinal data that have been published in robust samples, especially for modern versions of these tests. The lack of available robust longitudinal data for some modern tests (e.g., WAIS-IV) could affect validity when interpreting changes in test scores without access to appropriate data. Although the tests used in this study may be older versions, they should not necessarily be considered obsolete due to the fact that they are being used in large, modern, federally funded research projects on cognitive aging and neurodegenerative disease (e.g., the NACC UDS). In fact, one could argue that the availability of robust longitudinal data make these tests more appropriate than updated versions for serial assessment of older adults, especially if one takes the perspective that research evidence, and not test publishers, should dictate the selection of tests and test norms used by neuropsychologists (Adams, 2000; Bush, 2010; Silverstein & Nelson, 2000; Strauss, Spreen, & Hunter, 2000).

Many of the UDS neuropsychological tests have marginal test–retest reliability for measuring change in cognition across approximately annual evaluations (Strauss, Sherman, & Spreen, 2006). Although the lengthy interval between baseline and follow-up testing (M = 14.62 months; SD = 5.20) would be expected to cause a decrease in test–retest reliability relative to shorter intervals, these reliability data are thought to possess better external validity than reliability coefficients obtained at shorter intervals because approximately 1 year is believed to be a typical (or even shorter than typical) retest interval for older adults who are cognitively healthy at baseline. Because of these undesirable test–retest reliability values, the margin of error required to detect reliable change can be quite large for some tests (Table 4). Although this margin of error may not be sufficiently precise to detect subtle changes, these results may nevertheless be valuable for detecting more obvious cognitive decline across an approximately 1-year period. The results presented here suggest that there may be great value in focusing on test–retest reliability in the development of new cognitive tests, but interpretation of score changes must also account for demographic variables, past exposure to tests, and test–retest intervals.

https://doi.org/10.1017/S1355617715000582 Published online by Cambridge University Press

Table 6. Neuropsychological and RCI data for sample case

Test	Time 1	Percentile	Time 2	Percentile	Lower 90% RCI	Upper 90% RCI	Change *z*-score	Change BR
MMSE	29	58	27	8	27.2	30.9	−1.80	10.2%
DS-F	9	55	8	35	6.5	11.5	−0.66	35.8%
DS-B	7	53	8	71	4.4	9.9	0.51	81.4%
DSC	31	9	28	5	24.2	44.4	−1.03	27.8%
TMT-A	44	24	41	32	20.4	54.2	−0.36	66.4%
TMT-B	150	8	160	5	72.3	175.3	−1.16	27.9%
LM-I	11	27	11	27	8.0	17.7	−0.62	49.6%
LM-D	9	26	9	27	6.5	16.5	−0.81	47.7%
Animals	19	40	13	8	13.1	26.6	−1.68	11.3%
Vegetables	14	61	8	11	8.4	19.2	−1.77	6.4%
BNT	28	53	25	18	25.2	31.5	−1.74	4.5%

Note. RCI = Reliable Change Interval; BR = Base Rate; MMSE = Mini-Mental State Examination; DS-F = Digit Span Forward; DS-B = Digit Span Backward; DSC = Digit Symbol Coding; TMT = Trail Making Test; LM-I = Logical Memory Immediate Recall; LM-D = Logical Memory Delayed Recall; BNT = Boston Naming Test.

ACKNOWLEDGMENTS

The authors thank Stephen Hawes, Ph.D., and the rest of the NACC Publication Review Committee for their helpful comments on a previous draft of this manuscript. The NACC database is funded by NIA Grant U01 AG016976. The authors have no conflicts of interests to disclose.

REFERENCES

Adams, K.M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment*, *12*, 281–286.

Attix, D.K., Story, T.J., Chelune, G.J., Ball, J.D., Stutts, M.L., Hart, R.P., & Barth, J.T. (2009). The prediction of change: Normative neuropsychological trajectories. *The Clinical Neuropsychologist*, *23*, 21–38.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8 [Software]. Retrieved from http://CRAN.R-project.org/package=lme4.

Beekly, D.L., Ramos, E.M., Lee, W.W., Deitrich, W.D., Jacka M.E., Wu, J., … Kukull, W.A. (2007). The National Alzheimer's Coordinating Center (NACC) database: The Uniform Data Set. *Alzheimer Disease and Associated Disorders*, *21*, 249–258.

Bläsi, S., Zehnder, A.E., Berres, M., Taylor, K.I., Spiegel, R., & Monsch, A.U. (2009). Norms for change in episodic memory as a prerequisite for the diagnosis of mild cognitive impairment (MCI). *Neuropsychology*, *23*, 189–200.

Bush, S.S. (2010). Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations. *The Clinical Neuropsychologist*, *24*, 7–16.

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

De Santi, S., Pirraglia, E., Barr, W., Babb, J., Williams, S., Rogers, K., … de Leon, M.J. (2008). Robust and conventional neuropsychological norms: Diagnosis and prediction of age-related cognitive decline. *Neuropsychology*, *22*, 469–484.

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

Duff, K., Callister, C., Dennett, K., & Tometich, D. (2012). Practice effects: A unique cognitive variable. *The Clinical Neuropsychologist*, *26*, 1117–1127.

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

Frerichs, R.J., & Tuokko, H.A. (2005). A comparison of methods for measuring cognitive change in older adults. *Archives of Clinical Neuropsychology*, *20*, 321–333.

Heaton, R.K., Miller, S.W., Taylor, S.J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Lutz, FL: Psychological Assessment Resources, Inc.

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*, 1267–1278.

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*, 244–256.

Holtzer, R., Goldin, Y., Zimmerman, M., Katz, M., Buschke, H., & Lipton, R. B. (2008). Robust norms for selected neuropsychological tests in older adults. *Archives of Clinical Neuropsychology*, *23*, 531–541.

Jefferson, A.L., Wong, S., Gracer, T.S., Ozonoff, A., Green, R.C., & Stern, R.A. (2007). Geriatric performance on an abbreviated version of the Boston Naming Test. *Applied Neuropsychology*, *14*, 215–223.

Matarazzo, J.D., & Herman, D.O. (1984). Base rate data for the WAIS-R: Test-retest stability and VIQ-PIQ differences. *Journal of Clinical Neuropsychology*, *6*, 351–366.

McCaffrey, R.J., Duff, K., & Westervelt, H.J. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments*. New York, NY: Springer.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., … Phelps, C.H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from

the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*, 263–269.

Mitrushina, M., Boone, K.B., Razani, J., & D'Elia, L.F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Morris, J.C., Weintraub, S., Chui, H.C., Cummings, J., Decarli, C., Ferris, S., … Kukull, W.A. (2006). The Uniform Data Set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease and Associated Disorders*, *20*, 210–216.

Pinhiero, J.C., & Bates, D.M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Pedraza, O., Lucas, J., Smith, G.E., Petersen, R.C., Graff-Radford, N.R., & Ivnik, R.J. (2010). Robust and expanded norms for the Dementia Rating Scale. *Archives of Clinical Neuropsychology*, *25*, 347–358.

R Core Team (2015). *R: A language and environment for statistical computing* (Version 3.1.2) [Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Reitan, R., & Wolfson, D. (1993). *The Halstead-Reitan neuropsychological test battery: Theory and clinical applications*. Tucson, AZ: Neuropsychology Press.

Salthouse, T.A., & Tucker-Drob, E.M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, *22*, 800–811.

Shirk, S.D., Mitchell, M.B., Shaughnessy, L.W., Sherman, J.C., Locascio, J.J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy*, *3*, 32.

Silverstein, M.L., & Nelson, L.D. (2000). Clinical and research implications of revising psychological tests. *Psychological Assessment*, *12*, 298–303.

Sosa-Ortiz, A.L., Acosta-Castillo, I., & Prince, M.J. (2012). Epidemiology of dementias and Alzheimer's disease. *Archives of Medical Research*, *43*, 600–608.

Strauss, E., Sherman, E.M.S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York, NY: Oxford University Press.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment*, *12*, 237–244.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.

Wechsler, D. (1987). *WMS-R: Wechsler Memory Scale-Revised*. New York: Psychological Corporation.

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N.R., Chui, H., … Morris, J.C. (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): The neuropsychologic test battery. *Alzheimer Disease and Associated Disorders*, *23*, 91–101.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B., & Freckleton, R.P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*, 1182–1189.