





## Research Article

# Quick-reference criteria for identifying multivariate cognitive change in older adults with mild cognitive impairment and dementia: An ADNI study

Jeremy G. Grant<sup>1,2</sup> , Amanda M. Wisinger<sup>2</sup>, Hilary F. Abel<sup>2</sup>, Jennifer M. Hunter<sup>2</sup>, Glenn E. Smith<sup>2</sup>  and the Alzheimer's Disease Neuroimaging Initiative<sup>1</sup>

<sup>1</sup>Department of Psychology, The Ohio State University, Columbus, OH, USA and <sup>2</sup>Department of Clinical and Health Psychology, University of Florida, Gainesville, FL, USA

### Abstract

**Objective:** To establish quick-reference criteria regarding the frequency of statistically rare changes in seven neuropsychological measures administered to older adults. **Method:** Data from 935 older adults examined over a two-year interval were obtained from the Alzheimer's Disease Neuroimaging Initiative. The sample included 401 cognitively normal older adults whose scores were used to determine the natural distribution of change scores for seven cognitive measures and to set change score thresholds corresponding to the 5<sup>th</sup> percentile. The number of test scores that exceeded these thresholds were counted for the cognitively normal group, as well as 381 individuals with mild cognitive impairment (MCI) and 153 individuals with dementia. Regression analyses examined whether the number of change scores predicted diagnostic group membership beyond demographic covariates. **Results:** Only 4.2% of cognitively normal participants obtained two or more change scores that fell below the 5<sup>th</sup> percentile of change scores, compared to 10.6% of the stable MCI participants and 38.6% of those who converted to dementia. After adjusting for age, gender, race/ethnicity, and premorbid estimates, the number of change scores below the 5<sup>th</sup> percentile significantly predicted diagnostic group membership. **Conclusions:** It was uncommon for older adults to have two or more change scores fall below the 5<sup>th</sup> percentile thresholds in a seven-test battery. Higher change counts may identify those showing atypical cognitive decline.

**Keywords:** Cognitive aging; cognitive dysfunction; neuropsychological tests; memory; Alzheimer's disease; longitudinal studies

(Received 3 November 2023; final revision 9 May 2024; accepted 5 June 2024; First Published online 25 November 2024)

### Introduction

A key purpose of neuropsychological assessment is to provide objective data regarding cognitive functioning to reliably and validly identify cognitive change over time. In the context of Alzheimer's disease and related dementias (ADRD), a crucial task is distinguishing between normal age-related cognitive decline and atypical cognitive decline associated with neurodegenerative disease (Chelune & Duff, 2019). Just as medical laboratory tests are repeated to evaluate changes in disease progression or response to intervention, serial neuropsychological assessment can provide valuable information about the current course and the future trajectory of cognitive changes in an individual patient.

A variety of statistical methods have been developed to identify changes in individual test performances, including (a) simple discrepancy score, (b) standard deviation methods, (c) reliable change methods, (d) regression-based methods, and (Duff, 2012). The simple discrepancy score is a straightforward method in which the difference between raw scores at two time points is compared to a normative database to determine how frequently a specific

discrepancy is observed in a particular population (Patton et al., 2005). In standard deviation methods, a difference in test performance between two evaluations that exceeds a particular standard deviation cut-off is characterized as a significant change in cognitive ability (Frerichs & Tuokko, 2005). In reliable change methods, an observed difference in test performance that exceeds the amount of change expected from measurement error or practice effects is characterized as a significant cognitive change (Chelune et al., 1993; Jacobson & Truax, 1991; Stein et al., 2010). In standard regression-based methods, an individual's baseline and follow-up scores are entered into a regression equation to determine whether the magnitude of the observed change in test performance exceeds the predicted variability in test performance based on a control sample (Hammers et al., 2022; McSweeney et al., 1993). Reliable change and regression-based methods are particularly useful because they provide estimates of the degree of measurement error influencing test-retest difference scores and then allow the examiner to infer the extent to which the examinee has experienced a statistically reliable change in performance (Brooks et al., 2016).

**Corresponding author:** Jeremy G. Grant; Email: [grant.866@osu.edu](mailto:grant.866@osu.edu)

**Cite this article:** Grant J.G., Wisinger A.M., Abel H.F., Hunter J.M., & Smith G.E. (2024) Quick-reference criteria for identifying multivariate cognitive change in older adults with mild cognitive impairment and dementia: An ADNI study. *Journal of the International Neuropsychological Society*, 30: 944–953, <https://doi.org/10.1017/S1355617724000407>

© The Author(s), 2024. Published by Cambridge University Press on behalf of International Neuropsychological Society. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

There is considerable debate about which approaches to identifying cognitive change best predict real-world function, and each approach presents some important limitations (Heilbrunner et al., 2010). First, sophisticated statistical procedures like reliable change methods and regression-based methods are rarely used in clinical settings; clinicians often simply examine differences in raw scores between two evaluations and rely on subjective judgments of clinical significance. A single reliable change or regression equation can help identify changes in an individual measure, but a clinician would need to input data into several different equations to examine all the measures in a battery (Cysique et al., 2011; Woods et al., 2006). Furthermore, manually inputting data into multiple equations is less favorable under typical clinical time constraints. Second, when the standard deviation methods are employed, there is wide variability in the thresholds used to denote significant changes (e.g.,  $\pm 1.0$ ,  $\pm 1.5$ , or  $\pm 2.0$  standard deviations). Third, the demographic characteristics that influence the normative distributions from which these standard deviations are based (e.g., age, gender, education, race, and ethnicity) differ widely between neuropsychological tests and can create interpretation issues when comparing multiple tests in a given battery (Merkley et al., 2022). Therefore, there is a need for statistical approaches to identifying atypical cognitive change that can be quickly applied in clinical settings.

The primary aim of the current study was to develop quick-reference criteria for identifying atypical cognitive change in older adults using a novel and easily accessible approach: examining the number of change scores in a test battery that corresponds to a statistically rare magnitude of change across multiple measures in cognitively normal older adults. A secondary aim was to examine the extent to which multivariate changes in cognitive performance predict diagnostic status after accounting for other variables that are commonly used in demographic normative adjustments for neuropsychological tests, including age, gender, education, and race/ethnicity.

## Method

### Participants

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). The primary goal of ADNI has been to examine the extent to which imaging, clinical, and neuropsychological measures predict the progression of Alzheimer's disease (AD). ADNI data has been collected across four phases to date; ADNI1 (begun in 2004), ADNI GO (begun in 2009), ADNI 2 (2011), and ADNI 3 (2016). Please see the ADNI website (<https://adni.loni.usc.edu>) for a thorough review of the participating institutions and study phases. ADNI was approved by the institutional review boards (IRBs) of all participating institutions. Written informed consent was obtained from study participants or their proxy, following the ethical standards set forth by the Declaration of Helsinki.

Eligible participants at the initial ADNI screening visit were fluent English or Spanish speakers ages 55 to 90, with at least six years of formal education and no prior history of acquired brain injury or psychiatric disease. Participants were diagnosed as normal (NL), mild cognitive impairment (MCI), or dementia according to ADNI's classification of diagnostic categories. NL participants had no abnormal memory complaints, clinical dementia rating (CDR) scores of 0, Mini-Mental State Exam (MMSE) scores between 24 and 30, and performed within the following ranges on the Wechsler Memory Scale-Revised

(WMS-R) Logical Memory II delayed paragraph recall subtest: raw scores greater than or equal to 9 for individuals with 16 or more years of education, greater than or equal to 5 for those with 8–15 years of education, and greater than or equal to 3 for those with 0–7 years of education. MCI participants had abnormal memory complaints (verified by a study partner) with intact functioning in activities of daily living, CDR scores of 0.5, MMSE scores between 24 and 30, and WMS-R Logical Memory II raw scores less than or equal to 8 for individuals with 16 or more years of education, less than or equal to 4 for those with 8–15 years of education, and less than or equal to 2 for those with 0–7 years of education. Participants were diagnosed with dementia if they had abnormal memory complaints and the same Logical Memory II score ranges as the MCI participants but also had CDR scores of 0.5 or 1.0, MMSE scores between 20 and 26 and met the NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984).

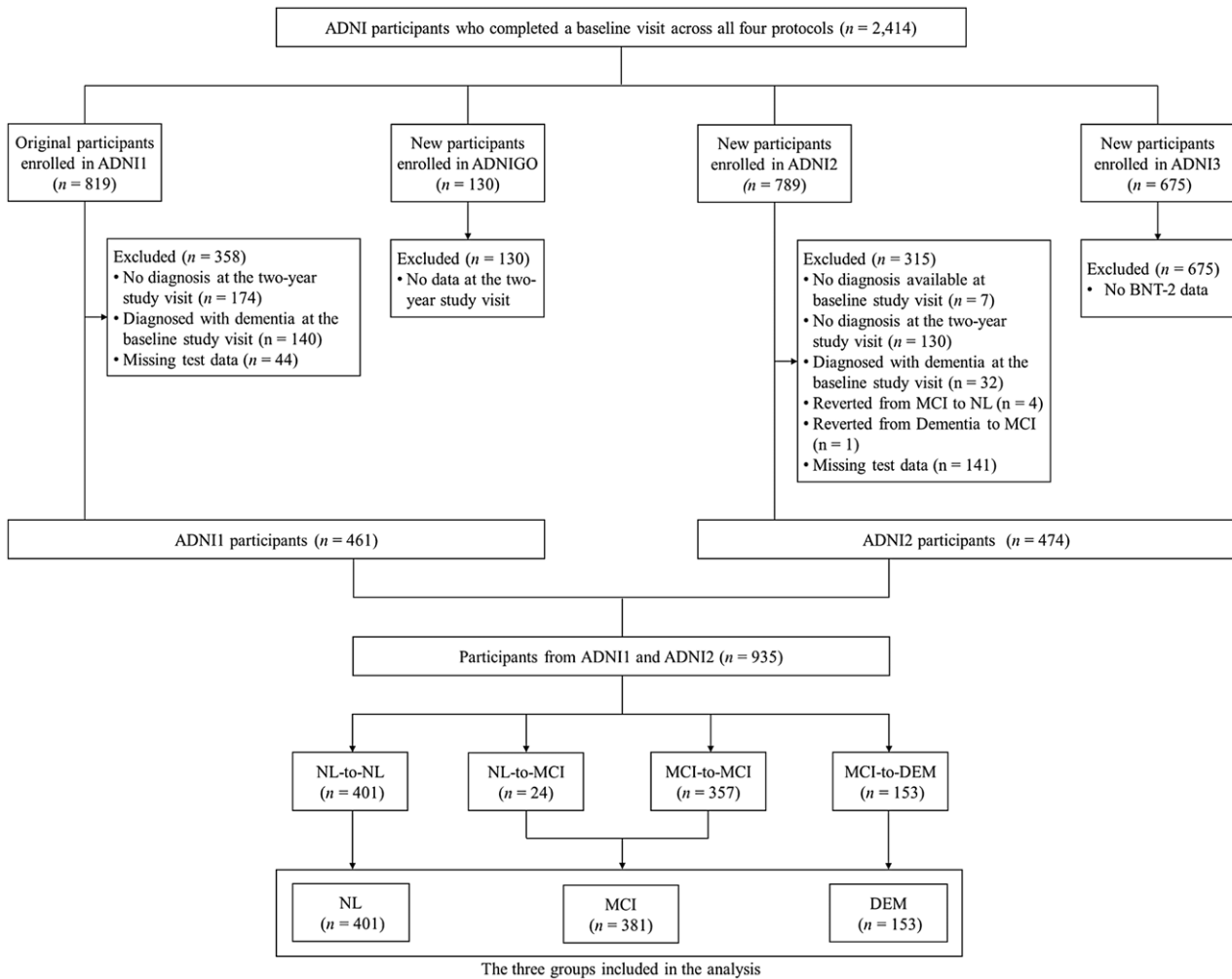
For the present study, inclusion criteria consisted of participants who (a) were diagnosed as cognitively normal or MCI at baseline, (b) were diagnosed as cognitively normal, MCI, or dementia at the two-year follow-up visit, and (c) had complete data on the following tests: American National Adult Reading Test (ANART; Grober et al., 1991); Boston Naming Test (BNT-2; Kaplan et al., 1983), Clock Drawing Test (Goodglass & Kaplan, 1983), Trail Making Test (Reitan, 1958); Category Fluency Test (Morris et al., 1989); and the Auditory Verbal Learning Test (RAVLT; Rey, 1964). Performance on the Digit Span Test and the second trial of the Category Fluency Test (vegetables) were also considered for this analysis; however, these two tasks were not administered in ADNI2 and were therefore not examined to maximize the diagnostic group sample sizes. Logical Memory II performance was not considered in defining the diagnostic groups.

As of July 17, 2022, a total of 2,414 participants had cognitive data across four ADNI protocols: ADNI1 (2004;  $n = 819$ ), ADNI2 (2011;  $n = 130$ ), ADNIGO (2009;  $n = 789$ ), and ADNI3 (2016;  $n = 675$ ). 358 ADNI1 participants were excluded due to consensus diagnostic classification of dementia at baseline, missing a diagnosis at the two-year study visit, or missing data on the cognitive measures of interest. 315 ADNI2 participants were excluded due to missing diagnoses at baseline or follow-up, a diagnosis of dementia at baseline, reversions in diagnosis (MCI to NL, or dementia to MCI), or missing test data. All participants from ADNIGO were excluded because they did not have cognitive data at the two-year mark, and all participants from ADNI3 were excluded due to unavailable data for the BNT-2, one of the measures of interest.

The final sample consisted of 935 participants from ADNI1 ( $n = 461$ ) and ADNI2 ( $n = 474$ ) who had complete and valid neuropsychological data at baseline and at the two-year study visit. Three groups were examined in this study based on diagnosis at their year 2 study visit: a cognitively normal group (NL;  $n = 401$ ), an MCI group (MCI;  $n = 381$ ), and a dementia group (DEM;  $n = 153$ ). The MCI group included 24 participants who had been diagnosed as normal at their baseline visit and 357 originally diagnosed as MCI. All participants had a baseline diagnosis of MCI. See Figure 1 for a visualization of the participant selection process for the current study.

### Statistical analyses

First, change scores were calculated using the difference between each participant's raw performance at baseline and at their two-year follow-up visit. Change scores were calculated for the



**Figure 1.** Flow diagram of ADNI participants included in the current study.

following seven measures: BNT-2 total correct (the sum of spontaneously correct responses and correct responses following stimulus cue); CDT total score (the sum of raw scores on command and copy); TMT Part A (TMT-A) and Part B (TMT-B); Animal fluency; the sum of the five immediate recall trials from the RAVLT (RAVLT Immediate) and the delayed recall trial (RAVLT Delayed). Change scores were calculated by subtracting the raw score at baseline from the raw score at follow-up. Second, the distribution of change scores in the NL cohort was examined for each of the seven measures. For each measure, the change score that corresponded to the 5<sup>th</sup> percentile of the distribution was identified as the threshold, denoting a significant decline in performance. Participants were classified as having exhibited a large change score on a given measure if their decline in performance met or exceeded the 5<sup>th</sup> percentile threshold (i.e., a worse decline in performance than 95% of the NL participants). Third, the number of change scores below the 5<sup>th</sup> percentile threshold was counted for each participant (henceforth called 5<sup>th</sup> percentile change count), thereby assigning a number ranging from 0 to 7 to each participant. Next, a stepwise multinomial logistic regression analysis was conducted to predict diagnostic group membership. Age, gender, race/ethnicity, years of education, and ANART error scores (as a measure of premorbid function) were entered in the initial model. ANART errors rather than

standard scores were used to minimize multicollinearity with education. Based on ADNI demographics, race/ethnicity was recoded into three categories: White (non-Hispanic), Black (non-Hispanic), and Other. The 5<sup>th</sup> percentile change count was entered in the following step as the primary variable of interest.

Finally, classification accuracy analyses were conducted to examine the sensitivity and specificity of a dichotomized 5<sup>th</sup> percentile change count for differentiating between the NL participants and those with atypical cognitive decline (i.e., the MCI and dementia participants).

In an exploratory analysis, the 10<sup>th</sup> percentile was also examined as a change score threshold for each measure to provide a less conservative estimate of cognitive decline. The results of the regression analyses using the 10<sup>th</sup> percentile change counts were nearly identical to those obtained with the 5<sup>th</sup> percentile change counts and are therefore not reported here.

## Results

### Sample characteristics

Table 1 displays the descriptive characteristics of each of the three diagnostic groups. Kruskal–Wallis H tests examined differences in age, years of education, and estimated premorbid intellect (NART education-corrected standard scores) between the diagnostic

**Table 1.** Demographic characteristics by diagnostic group

Variable	NL	MCI	Dementia
<i>n</i>	401	381	153
Age (years)	74.3 ± 5.8	72.7 ± 7.5	73.5 ± 7.1
Gender (% women)	50.1%	40.9%	40.5%
Race/Ethnicity			
White (Non-Hispanic/Latino)	91.5%	90.6%	96.7%
Black (Non-Hispanic/Latino)	4.0%	4.2%	0.7%
White (Hispanic/Latino)	2.0%	1.6%	1.3%
Mixed Race (Hispanic/Latino)	0.2%	0.3%	0.7%
Asian	1.5%	1.8%	0.7%
Native American/Alaskan Native	0.2%	–	–
Native Hawaiian/Pacific Islander	–	0.3%	–
More than one race/ethnicity	0.5%	1.3%	–
Education (years)	16.4 ± 2.7	16.1 ± 2.8	16.0 ± 2.7
Premorbid Intellect (ANART SS)	101.3 ± 7.7	98.8 ± 7.6	97.9 ± 8.0

Note: Only shows the statistically significant differences between diagnostic groups, obtained via Kruskal–Wallis H Tests.

groups. Distributions of age, years of education, and estimated premorbid intellect were similar for all groups, as assessed by visual inspection of boxplots. The diagnostic groups did not differ in years of education ( $H(2) = 3.628, p = .163$ ). However, the median age was statistically significantly different between groups ( $H(2) = 6.524, p = .038$ ); the MCI group was younger than the NL group but not significantly younger than the dementia group. In addition, the MCI and dementia groups showed lower premorbid intellect ( $H(2) = 36.280, p < .001$ ) than the NL group.

### Cognitive performance at baseline and follow-up

Table 2 displays cognitive performance at baseline and two-year follow-up for each diagnostic group. There appeared to be a slight practice effect for the NL group; these participants tended to perform slightly better in the follow-up visit across measures. However, there was less evidence of a practice effect for the other two groups; the MCI and dementia participants tended to perform at the same level or slightly worse in the follow-up visit.

### The distribution of change scores for the cognitively normal participants

Figure 2 displays the distribution of change scores for the Rey Auditory Verbal Learning Test (RAVLT) Delayed Recall trial. The median and modal change score on this task was zero; 16.7% of NL participants obtained a change score of 0 (i.e., recalled the exact same number of words on the delayed recall trial at baseline and follow-up). 48.4% of NL participants exhibited an improvement in RAVLT Delayed Recall performance at two-year follow-up (ranging from +1 to +15 words), and 34.9% exhibited a decline in delayed recall (ranging from –1 to –14 words). Of those who exhibited a decline, 20 participants (4.98% of the sample) obtained change scores less than or equal to –7 (i.e., they recalled seven words or more at baseline than they did at follow-up). Therefore, –7 served as the 5<sup>th</sup> percentile change score threshold for the RAVLT Delayed Recall subtest; individuals who obtained change scores of –7 or lower were classified as having exhibited a significant decline in performance. The change score distributions for the other six measures were examined similarly; as a second example, Figure 3 displays the change score distribution for the Category Fluency test.

Table 3 displays descriptive statistics for the change scores, as well as the change scores that corresponded to the 5<sup>th</sup> percentile threshold for each of the seven measures. Table 4 displays the

associations between the 5<sup>th</sup> percentile change count and relevant demographic characteristics for the overall sample. The 5<sup>th</sup> percentile change count was not associated with age, gender, education, race/ethnicity, or premorbid intellect.

### Comparing the number of substantial change scores across diagnostic groups

Figure 4 shows the 5<sup>th</sup> percentile change count across diagnostic groups for all seven cognitive measures. Among the NL participants, the grand majority (75.3%) did not have any significant change scores below the 5<sup>th</sup> percentile thresholds across seven measures, and one-fifth (20.4%) had only one significant change score. It was increasingly rare for the NL participants to have significant change scores across multiple variables; only 3.7% had a 5<sup>th</sup> percentile change count of two or more, and only 0.5% had a change count of three or more. By comparison, the MCI participants demonstrated a slightly higher proportion of large declines in performance, with 22.3% having a 5<sup>th</sup> percentile change count of one or more and 6.8% having two or more. The dementia group had the highest proportion of participants with significant change scores at every level, with 27.5% having a 5<sup>th</sup> percentile change count of one or more and 20.9% having two or more.

Figure 5 displays the cumulative percentage of participants in each diagnostic group with significant change scores. A 5<sup>th</sup> percentile change count of one or more was relatively common for all three groups. A 5<sup>th</sup> percentile change count of two or more was rare for the NL group (4.2%), relatively rare for the MCI group (10.6%), but relatively common for the dementia group (38.6%). A 5<sup>th</sup> percentile change count of three or more was relatively rare for all three groups.

### The predictive value of the number of change scores toward diagnostic status

Table 5 displays the parameter estimates for the logit predicting NL versus MCI group membership and the logit predicting NL versus dementia group membership. The NL group served as the reference category for pairwise comparisons of the odds ratio predicting MCI or dementia group membership. The fit between the stepwise multinomial logistic regression model containing age, gender, race/ethnicity, education, and premorbid intellect was significantly improved with the addition of the dichotomized 5<sup>th</sup> percentile change count ( $X^2(14) = 206.27, p < .001$ , Nagelkerke  $R^2 = .23$ ).

For the logit predicting NL versus MCI group membership, younger age ( $\text{Exp}(\beta) = .96, p < .001$ ), male gender ( $\text{Exp}(\beta) = .158, p = .003$ ), lower premorbid intellect ( $\text{Exp}(\beta) = 1.05, p < .001$ ), and the 5<sup>th</sup> percentile change count ( $\text{Exp}(\beta) = 1.57, p < .001$ ) were all associated with greater odds of MCI group membership. Education and race/ethnicity did not significantly predict NL versus MCI group membership. For the logit predicting NL versus dementia group membership, male gender ( $\text{Exp}(\beta) = 1.63, p = .029$ ), lower premorbid intellect ( $\text{Exp}(\beta) = 1.07, p < .001$ ), and the 5<sup>th</sup> percentile change count ( $\text{Exp}(\beta) = 3.48, p < .001$ ) were associated with greater odds of dementia. In contrast, Black participants ( $\text{Exp}(\beta) = .08, p = .018$ ) had lower odds of dementia relative to NL group membership. Age and education did not significantly predict NL versus dementia group membership.

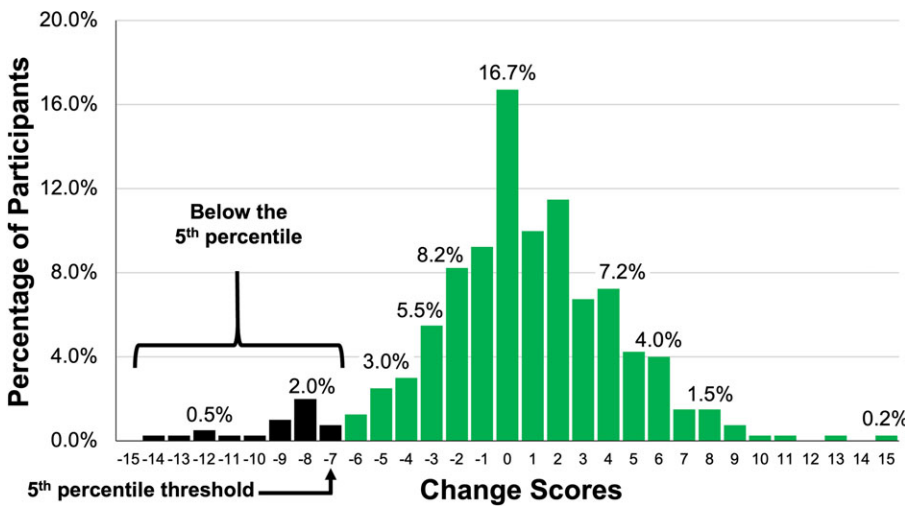
Table 6 displays the classification matrix differentiating between the diagnostic groups for using the 5<sup>th</sup> percentile change count. For this analysis, participants with MCI and dementia were combined into a single group representing all individuals with atypical cognitive decline (i.e., a “positive” diagnosis) in



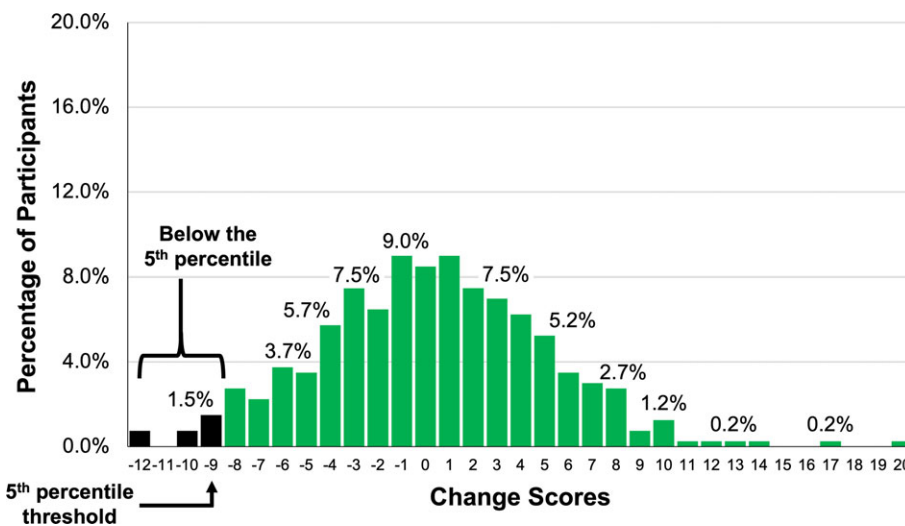
**Table 2.** Cognitive performance (raw scores) at baseline and 2-year follow-up visit, by diagnostic group

Cognitive Measure	NL	MCI	Dementia
<i>n</i>	401	381	153
Clock Drawing Test (baseline)	9.6 ± 0.8	9.2 ± 1.1	8.5 ± 1.5
Clock Drawing Test (follow-up)	9.5 ± 0.9	9.2 ± 1.2	8.2 ± 1.9
Boston Naming Test (baseline)	28.3 ± 2.0	26.7 ± 3.4	25.5 ± 3.9
Boston Naming Test (follow-up)	28.5 ± 2.0	26.9 ± 3.9	24.2 ± 5.1
Category Fluency (baseline)	20.9 ± 5.4	17.7 ± 5.1	15.6 ± 4.5
Category Fluency (follow-up)	21.1 ± 5.4	17.1 ± 5.3	12.4 ± 5.1
Trail Making Test-Part A (baseline)	33.9 ± 11.4	37.9 ± 15.1	47.5 ± 23.2
Trail Making Test-Part A (follow-up)	31.7 ± 10.9	39.0 ± 20.0	56.4 ± 29.9
Trail Making Test-Part B (baseline)	81.4 ± 38.1	102.2 ± 51.3	139.7 ± 73.3
Trail Making Test-Part B (follow-up)	80.2 ± 38.5	107.1 ± 58.9	191.0 ± 95.4
ALVT Immediate Recall (baseline)	45.3 ± 9.6	36.1 ± 10.6	28.1 ± 7.2
ALVT Immediate Recall (follow-up)	46.3 ± 10.4	35.3 ± 12.1	23.1 ± 8.1
RAVLT Delayed Recall (baseline)	7.7 ± 3.8	4.4 ± 4.0	1.4 ± 2.0
RAVLT Delayed Recall (follow-up)	8.2 ± 4.1	3.9 ± 4.3	0.4 ± 1.1

Note: Kruskal-Wallis H Tests with Bonferroni corrections to account for multiple comparisons.



**Figure 2.** Distribution of change scores on the RAVLT delayed recall subtest for the cognitively normal (NL) participants.



**Figure 3.** Distribution of change scores on the category fluency test for the cognitively normal (NL) group.

comparison to the NL participants (i.e., a “negative” diagnosis). In addition, the 5<sup>th</sup> percentile change count was dichotomized such that participants with two or more change scores below the 5<sup>th</sup> percentile served as the “positive” test result, and participants with

zero or only one change score below the 5<sup>th</sup> percentile served as the “negative” test result. The dichotomized 5<sup>th</sup> percentile change count showed high specificity (96%) and high positive predictive value (85%) for differentiating between the diagnostic groups but

**Table 3.** Thresholds corresponding to the 5th percentile in the distribution of change scores for the cognitively normal (NL) group

Cognitive Measure	Mean Change Score $\pm$ SD	Median Change Score	Greatest Improvement in Performance	Worst Decline in Performance	5 <sup>th</sup> percentile threshold
Clock Drawing Test	-0.08 $\pm$ 0.96	0	+4	-4	$\leq$ -3
Boston Naming Test	0.28 $\pm$ 1.51	0	+7	-6	$\leq$ -3
Category Fluency	0.26 $\pm$ 4.79	0	+20	-12	$\leq$ -9
Trail Making Test – Part A	-2.18 $\pm$ 11.71	-2	-88	+58	$\geq$ +14
Trail Making Test – Part B	-1.23 $\pm$ 38.51	-1	-230	+250	$\geq$ +53
RAVLT Immediate Recall	1.03 $\pm$ 8.23	+1	+32	-30	$\leq$ -14
RAVLT Delayed Recall	0.46 $\pm$ 3.94	0	+15	-14	$\leq$ -7

Note: Higher scores on the Trail Making Test indicate worse performance (i.e., a longer time to complete a timed task). Therefore, higher scores at 2-year follow-up indicate a decline in performance.

**Table 4.** The association between 5<sup>th</sup> percentile change count and demographic characteristics

Spearman correlations	5 <sup>th</sup> Percentile change count
Age	.02
Gender	.01
Education (years)	-.04
Race/ethnicity	-.04
Premorbid intellect (ANART errors)	.03

Note: Point-biserial correlations were used for the association between 5<sup>th</sup> Percentile Change Count and gender. \*  $p = 0.01$ .

low negative predictive value (47%) and low sensitivity (19%). The analysis thereby confirmed the findings illustrated in Figure 5; a 5<sup>th</sup> percentile change count of two or more might distinguish individuals with MCI or dementia from cognitively normal individuals, whereas a 5<sup>th</sup> percentile change count of zero or one does not distinguish well between the diagnostic groups.

## Discussion

This study aimed to provide proof-of-concept for a novel, quickly-accessible method for identifying meaningful changes in neuropsychological test performances over time. By examining the performances of a large sample of older adults diagnosed as cognitively intact, calculating their change scores across a two-year interval, and establishing the magnitude of change needed to be considered normatively rare for each measure, this method could allow clinicians to estimate whether an examinee's performances are atypical in comparison to other older adults who present as clinically normal. Participants diagnosed as cognitively normal at baseline and at two-year follow-up served as the normative reference group for establishing the criteria by which abnormally large declines in performance were identified; these criteria were then validated in participants diagnosed with MCI at both time points, as well as in those who transitioned from cognitively normal to MCI or from MCI to dementia. Establishing base rates of multivariate cognitive change may help improve the value of neuropsychological evaluations in the diagnosis of neurological disease (Donders, 2020; Jak et al., 2009).

It was relatively common for participants to show a substantial decline on at least one of the seven cognitive measures in this study, regardless of diagnostic group. When assessed at baseline and two years later, roughly one-quarter of the participants diagnosed as cognitively normal at both time points showed one or more declines in performance that fell below the cut-off score

corresponding to the 5<sup>th</sup> percentile in the distribution of change scores for each measure. Similarly, it was also common for the MCI participants (one-third of the sample) to have at least one change score below the 5<sup>th</sup> percentile threshold. Among the participants who converted from MCI to dementia, exhibiting a substantial decline in cognitive performance was the rule rather than the exception; it was more common for these individuals to exhibit at least one large decline in performance (two-thirds of the group) across the seven measures than it was to obtain a change score above the 5<sup>th</sup> percentile thresholds. These findings lend further support to an established body of research demonstrating that it is common for cognitively intact and cognitively impaired individuals to exhibit at least one large change in performance over time when examining multiple neuropsychological tests. This phenomenon has been repeatedly shown in research employing reliable change methods (Binder et al., 2009; Brooks et al., 2016), highlighting the dangers of overinterpreting a decline in performance on a single measure.

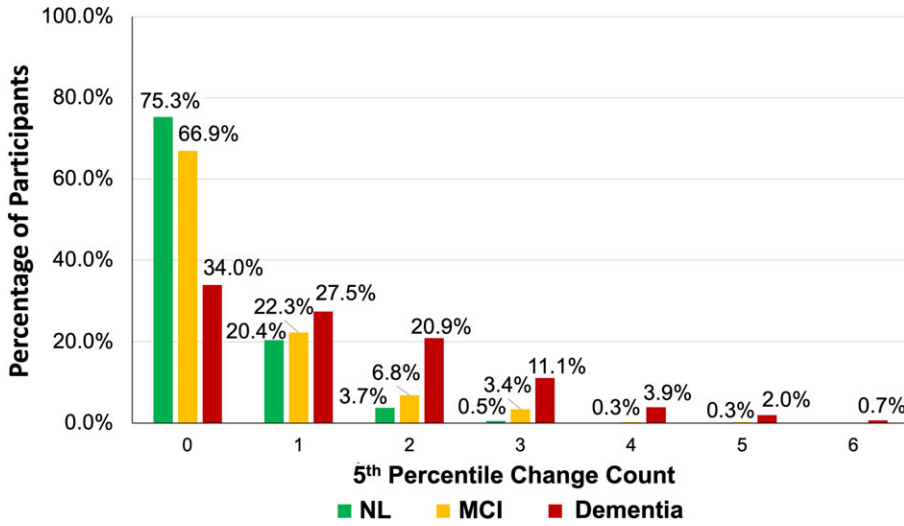
Between-group differences in the 5<sup>th</sup> percentile change count became more apparent as the criterion moved from having one or more large declines in performance to having two or more large declines. It remained relatively uncommon for cognitively normal participants to have a 5<sup>th</sup> percentile change count of two or more (less than 5% of the group). The overwhelming majority of the cognitively normal participants (over 95%) either had no declines or only one decline that exceeded the 5<sup>th</sup> percentile threshold for each of the seven measures. It was slightly more common for the MCI participants (about one-tenth of the group) to have a 5<sup>th</sup> percentile change count of two or more. Importantly, a large majority of the MCI participants (the remaining nine-tenths) still had no declines or only one decline in performance that fell below the 5<sup>th</sup> percentile thresholds, similar to the cognitively normal participants. The most notable between-group differences emerged when examining participants who converted from MCI to dementia. In this group, it was relatively common (over one-third) to exhibit at least two large declines in performance, a much larger proportion than the other MCI and cognitively normal participants. These results indicate that having two or more large declines may be a useful criterion for distinguishing between the typical variability seen in clinically normal individuals and atypical cognitive decline.

One aspect of the study that warrants further investigation is the classification accuracy of the 5<sup>th</sup> percentile change count, particularly its low sensitivity for identifying participants in the two cognitively impaired groups. Classifying participants based on a 5<sup>th</sup> percentile change count of two or more yielded a high false negative rate: most of the participants in the combined MCI and

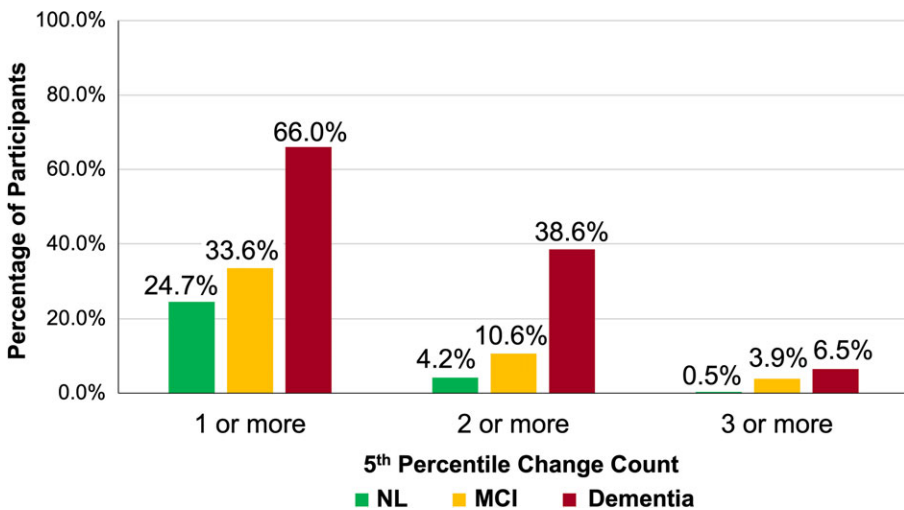
**Table 5.** Stepwise multinomial logistic regression predicting diagnostic group with 5<sup>th</sup> percentile threshold

Variables predicting NL vs. MCI group membership	B	Odds ratio	p
<i>Step 1</i>			
Age (years)	-.04	.96	<.001
Gender (Male vs. Female)	.46	1.58	.003
Race (White vs. Black)	-.26	.77	.505
Race (White vs. Other race/ethnicity)	.02	1.02	.957
Education (years)	-.02	.98	.464
Premorbid Intellect (ANART errors)	.05	1.05	<.001
<i>Step 2</i>			
5 <sup>th</sup> Percentile Change Count (zero or one vs. two or more)	.45	1.57	<.001
Variables Predicting NL vs. Dementia Group Membership	B	Odds Ratio	p
<i>Step 1</i>			
Age (years)	-.03	.98	.112
Gender (Male vs. Female)	.49	1.63	.029
Race (White vs. Black)	-2.52	.08	.018
Race (White vs. Other race/ethnicity)	-.84	.43	.176
Education (years)	-.04	.96	.327
Premorbid Intellect (ANART errors)	.068	1.07	<.001
<i>Step 2</i>			
5 <sup>th</sup> Percentile Change Count (zero or one vs. two or more)	1.247	3.48	<.001

Note: For ANART errors, a higher number of errors indicates worse performance. Therefore, an odds ratio greater than 1.00 indicates a greater likelihood of lower premorbid intellect.



**Figure 4.** Percentage of participants with significant change scores by diagnostic group. Note. NL = Cognitively normal; MCI = Mild cognitive impairment. Note: 5<sup>th</sup> Percentile Change Count = the number of significant change scores below the 5<sup>th</sup> percentile in the natural distribution of change scores for the NL group.



**Figure 5.** Cumulative percentage of participants with significant change scores by diagnostic group. Note. NL = Cognitively normal; MCI = Mild cognitive impairment. 5<sup>th</sup> Percentile Change Count = the number of significant change scores, below the 5<sup>th</sup> percentile in the natural distribution of change scores for the NL group. Participants with a y of “1 or more” includes those with 1, 2, 3, 4, 5, or 6 significant change scores below the 5<sup>th</sup> percentile.

**Table 6.** Classification matrix: using a dichotomized 5<sup>th</sup> percentile change count to differentiate between diagnostic groups

Test result		Diagnosis		Positive Predictive Value = 85% Negative Predictive Value = 47%
		+	-	
+	5 <sup>th</sup> Percentile Change Count = Two or More	MCI + Dementia True Positives = 100	NL False Positives = 17	
	5 <sup>th</sup> Percentile Change Count = Zero or One	False Negatives = 434	True Negatives = 384	
		Sensitivity = 19%	Specificity = 96%	

dementia group obtained only one or no large change scores exceeding the 5<sup>th</sup> percentile thresholds. One possible contributor to the low sensitivity of the change count metric may be floor effects. For example, on the RAVLT Delayed Recall test, the average baseline recall was seven words in the cognitively normal group, whereas, in the MCI group, the average baseline recall was just over four words for the MCI group; in the group who converted to dementia, the average baseline recall was less than two words. The 5<sup>th</sup> percentile threshold for the RAVLT Delayed Recall test corresponded to recalling seven fewer words at follow-up than at baseline. Thus, a substantial portion of the participants in the two cognitively impaired groups were not able to exhibit large enough declines to exceed this threshold. Therefore, a 5<sup>th</sup> percentile change count of two or more measures may have missed many participants who actually performed substantially worse at follow-up but remained within the threshold because they were too near the floor, thereby contributing to the high false negative rate. Floor effects may limit the use of the 5<sup>th</sup> percentile change count in clinical settings. If an examinee's score reaches the floor in a follow-up evaluation, a clinician could interpret their poor performance as a substantial decline by clinical judgment, without relying on base rates for that measure. However, if the examinee's score is well above the floor, the 5<sup>th</sup> percentile threshold could be useful for distinguishing between typical and atypical cognitive change across an entire battery.

Although the 5<sup>th</sup> percentile change count demonstrated low sensitivity, it is important to note that it yielded a high positive predictive value (PPV), which is the more clinically relevant metric (Smith et al., 2008). Because it was rare for cognitively normal individuals to show two or more large declines in performance across seven neuropsychological measures, obtaining such a score has PPV for non-normality. Furthermore, the data collection procedures used in ADNI intentionally oversampled cognitively normal individuals (relative to a clinical setting). Given the characteristics of individuals referred to undergo clinical neuropsychological evaluations, it is likely that the base rate of atypical cognitive decline will be higher in clinical settings than in research settings. And because positive (and negative) predictive values are strongly influenced by base rates, it is likely that the 5<sup>th</sup> percentile change count of two or greater would have an even larger PPV in the clinical scenario. Of course, a typical neuropsychological battery involves more than seven measures for which change scores could be calculated. Future research should explore how the 5<sup>th</sup> percentile change count changes with battery size. Nevertheless, these findings provide proof of concept that examining raw change scores in a large neuropsychological database could generate quickly accessible information. This information could be used to guide expectations about typical versus atypical cognitive change that can be applied to individual patients.

### Strengths, limitations, and directions for future research

A strength of this novel method is the ability to examine multivariate cognitive change. In contrast to reliable change and regression-based methods, in which each individual change score can only be examined in isolation, this novel, whole-battery approach allows examiners to make inferences about changes across multiple measures at once. Although some prior research has examined multivariate approaches for assessing meaningful cognitive change (Cysique et al., 2011; Woods et al., 2006), these studies rely on modified reliable change methods and/or standard regression-based approaches that remain underused by clinicians. A unique aspect of this study is using multivariate cognitive change to distinguish between diagnostic groups; prior research on this topic has typically been restricted to cognitively normal samples and has lacked external clinical samples that would help validate findings (Woods et al., 2006). Another strength of the current study is the lack of confounding variables. For example, the finding that individuals with stable MCI obtained fewer significant change scores than those who transitioned to dementia did not simply occur because cognitive performance was part of the diagnostic criteria to define the groups in the first place (i.e., performance on WMS-III Logical Memory II). The findings indicate that multivariate changes in cognitive performance across multiple cognitive domains may add value towards predicting diagnostic status in older adults.

The present study has several important limitations that influence the generalizability of the findings. First, the utility of this novel approach to identifying significant multivariate change is restricted by the limited neuropsychological test battery examined in this study. Presumably, larger multivariable batteries would produce a greater number of change scores that would normatively fall below a 5<sup>th</sup> percentile threshold. Additionally, a two-year interval between baseline and follow-up evaluations was selected to maximize the number of available data points for participants converting to MCI or dementia. The change score thresholds demonstrated in this study are less generalizable to neuropsychological evaluations performed over shorter or longer time periods. A second important limitation revolves around the population sampled. The present study was specifically designed to assist with the detection of abnormal cognitive aging, using a large sample of older adult research participants in ADNI as the normative reference. Therefore, the change score thresholds established in this study cannot be validly used to examine multivariate cognitive change in other populations where detecting cognitive change is of great interest, including healthy pediatric and adult populations as well as neurological populations where a gradual decline is not necessarily the expected cognitive trajectory (e.g., post-surgical epilepsy, brain injury populations). Likewise, the predominantly non-Hispanic White sample across all of the ADNI protocols



currently available hinders the generalizability of the findings to other racial/ethnic groups (Mindt et al., 2022). This is particularly relevant for detecting meaningful cognitive change in historically marginalized racial/ethnic groups who are at heightened risk for cognitive decline.

Each of the aforementioned limitations pose an opportunity to promote this field of research in the future. The study should be replicated in a large, ethnically diverse sample to examine base rates of substantial declines in performance using a larger battery that more closely resembles the number of measures obtained in a typical neuropsychological evaluation. A larger test battery would likely yield a larger 5<sup>th</sup> percentile change count necessary to identify atypical cognitive change. This study focused on examining cognitive decline rather than increases in cognitive performance; future research should explore the extent to which this approach can be applied to neuromedically stable populations (Cysique et al., 2011) and other populations where gradual improvement is a likely cognitive trajectory (e.g., mild traumatic brain injury). Future studies should also examine whether specific cognitive domains within a test battery can help distinguish between diagnostic groups. Previous work on reliable change in cognitively normal older adults has focused on memory (Binder et al., 2009; Brooks et al., 2007); the extent to which declines in other cognitive domains offer unique insights into typical and atypical cognitive change should be explored.

### Conclusion

This study demonstrates how examining the multivariate distribution of change scores among cognitively normal older adults may provide normative information for identifying atypical cognitive change. Among older adults assessed over a two-year interval, it was statistically rare to have two or more change scores out of seven measures that fell below the 5<sup>th</sup> percentile in the distribution of change scores. Older adults who exhibit multivariate changes in performance that exceed these standards are likely experiencing atypical cognitive decline. More research is needed to validate this simple method of examining multivariate cognitive change.

**Acknowledgments.** This work was supported in part by the Florida Department of Health, Public Health Research, Biomedical Research Program.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

**Competing interests.** The authors have no conflicts of interest to declare.

### References

- Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, 24(1), 31–46. <https://doi.org/10.1093/arclin/acn001>
- 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. <https://doi.org/10.1093/arclin/acw079>
- Brooks, B. L., Iverson, G. L., & White, T. (2007). Substantial risk of "Accidental MCI" in healthy older adults: Base rates of low memory scores in neuropsychological assessment. *Journal of the International Neuropsychological Society*, 13(3), 490–500. <https://doi.org/10.1017/S1355617707070531>
- Chelune, G. J., & Duff, K. (2019). The assessment of change: serial assessments in dementia evaluations. In L. D. Ravdin, & H. L. Katzen (Eds.), *Handbook on the neuropsychology of aging and dementia* (pp. 61–76). Springer International Publishing, [https://doi.org/10.1007/978-3-319-93497-6\\_5](https://doi.org/10.1007/978-3-319-93497-6_5)
- Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, 7(1).
- Cysique, L. A., D. Franklin, Abramson, I., Ellis, R. J., Letendre, S., Collier, A., Clifford, D., Gelman, B., McArthur, J., Morgello, S., Simpson, D., McCutchan, J. A., Grant, I., Heaton, R. K., the CHARTER group, the HNRC group (2011). Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. *Journal of Clinical and Experimental Neuropsychology*, 33(5), 505–522. <https://doi.org/10.1080/13803395.2010.535504>
- Donders, J. (2020). The incremental value of neuropsychological assessment: A critical review. *Clinical Neuropsychologist*, 34(1), 56–87. <https://doi.org/10.1080/13854046.2019.1575471>
- Duff, K. (2012). Current topics in science and practice evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, 27(3), 248–261. <https://doi.org/10.1093/arclin/acr120>
- Frerichs, R. J., & Tuokko, H. A. (2005). A comparison of methods for measuring cognitive change in older adults. *Archives of Clinical Neuropsychology*, 20(3), 321–333. <https://doi.org/10.1016/j.acn.2004.08.002>
- Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders*. Cambridge University Press.
- Grober, E., Sliwinski, M., & Korey, S. R. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, 13(6), 933–949. <https://doi.org/10.1080/01688639108405109>
- Hammers, D. B., Kostadinova, R., Unverzagt, F. W., & Apostolova, L. G. (2022). Assessing and validating reliable change across ADNI protocols. *Journal of Clinical and Experimental Neuropsychology*, 44(2), 85–102. <https://doi.org/10.1080/13803395.2022.2082386>
- 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. *Clinical Neuropsychologist*, 24(8), 1267–1278. <https://doi.org/10.1080/13854046.2010.526785>
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to denning meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12–19.
- Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., & Delis, D. C. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *American Journal of Geriatric Psychiatry*, 17(5), 368–375. <https://doi.org/10.1097/JGP.0b013e31819431d5>
- Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston naming test*. Philadelphia: Lea & Fibiger.

- McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease. *Neurology*, 34(7), 939–939. <https://doi.org/10.1212/WNL.34.7.939>
- McSweeney, A. J., Naugle, R. I., Chelune, G. J., & Lüders, H. (1993). "TScores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist*, 7(3), 300–312. <https://doi.org/10.1080/13854049308401901>
- Merkley, T. L., Esopenko, C., Zizak, V. S., Bilder, R. M., Strutt, A. M., Tate, D. F., & Irimia, A. (2022). Challenges and opportunities for harmonization of cross-cultural neuropsychological data. *Neuropsychology*, 37(3), 237–246. <https://doi.org/10.1037/neu0000818>
- Mindt, M. R., Okonkwo, O., Weiner, M. W., Veitch, D. P., Aisen, P., Ashford, M., Coker, G., Donohue, M. C., Langa, K. M., Miller, G., Petersen, R., Raman, R., Nosheny, R. (2022). Improving generalizability and study design of Alzheimer's disease cohort studies in the United States by including under-represented populations. *Alzheimer's and Dementia*, 19(4), 1549–1557. <https://doi.org/10.1002/alz.12823>
- Moms, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., Mellits, E. D., Clark, C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, 39(9), 1159–1159. <https://doi.org/10.1212/WNL.39.9.1159>
- Patton, D. E., Duff, K., Schoenberg, M. R., Mold, J., Scott, J. G., & Adams, R. L. (2005). Base rates of longitudinal RBANS discrepancies at one- and two-year intervals in community-dwelling older adults. *The Clinical Neuropsychologist*, 19(1), 27–44. <https://doi.org/10.1080/13854040490888477>
- Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8(3), 271–276. <https://doi.org/10.2466/pms.1958.8.3.271>
- Rey, A. (1964). *L'examen clinique en psychologie (The Clinical Psychological Examination)*. Presses Universitaires De France.
- Smith, G. E., Ivnik, R. J., & Lucas, J. A. (2008). Assessment techniques: tests, test batteries, norms and methodological approaches. In J. Morgan, & J. Ricker (Eds.), *Textbook of clinical neuropsychology*. Taylor & Francis.
- Stein, J., Luppá, M., Brähler, E., König, H.-H., & Riedel-Heller, S. G. (2010). The assessment of changes in cognitive functioning: Reliable change indices for neuropsychological instruments in the elderly – a systematic review. *Dementia and Geriatric Cognitive Disorders*, 29(3), 275–286. <https://doi.org/10.1159/000289779>
- Woods, S. P., Childers, M., Ellis, R. J., Guaman, S., Grant, I., Heaton, R. K., & HIV Neurobehavioral Research Center (HNRC) Group (2006). A battery approach for measuring neuropsychological change. *Archives of Clinical Neuropsychology*, 21(1), 83–89. <https://doi.org/10.1016/j.acn.2005.07.008>