




Discrepancy-Based Evidence for Loss of Thinking Abilities (DELTA): Development and Validation of a Novel Approach to Identifying Cognitive Changes

Breton M. Asken^{1,2,*} , Kelsey R. Thomas^{3,4} , Athene Lee^{1,5} , Jennifer D. Davis^{1,6}, Paul F. Malloy^{1,5}, Stephen P. Salloway^{1,5} and Stephen Correia for the Alzheimer's Disease Neuroimaging Initiative^{1,7,†}

¹Department of Psychiatry and Human Behavior, Alpert Medical School of Brown University, Providence, RI 02906, USA

²Department of Clinical and Health Psychology, University of Florida, Gainesville, FL 32610, USA

³Research Service, Veterans Affairs San Diego Healthcare System, San Diego, CA 92161, USA

⁴Department of Psychiatry, University of California, San Diego School of Medicine, La Jolla, CA 92093, USA

⁵Butler Hospital, Memory and Aging Program, Providence, RI 02906, USA

⁶Department of Psychiatry, Rhode Island Hospital, Providence, RI 02905, USA

⁷Mental Health and Behavioral Science Service, Providence VA Medical Center, Providence, RI 02908, USA

(RECEIVED June 28, 2019; FINAL REVISION September 25, 2019; ACCEPTED October 29, 2019; FIRST PUBLISHED ONLINE December 11, 2019)

Abstract

Objective: To develop and validate the Discrepancy-based Evidence for Loss of Thinking Abilities (DELTA) score. The DELTA score characterizes the strength of evidence for cognitive decline on a continuous spectrum using well-established psychometric principles for improving detection of cognitive changes. **Methods:** DELTA score development used neuropsychological test scores from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (two tests each from Memory, Executive Function, and Language domains). We derived regression-based normative reference scores using age, gender, years of education, and word-reading ability from robust cognitively normal ADNI participants. Discrepancies between predicted and observed scores were used for calculating the DELTA score (range 0–15). We validated DELTA scores primarily against longitudinal Clinical Dementia Rating-Sum of Boxes (CDR-SOB) and Functional Activities Questionnaire (FAQ) scores (baseline assessment through Year 3) using linear mixed models and secondarily against cross-sectional Alzheimer's biomarkers. **Results:** There were 1359 ADNI participants with calculable baseline DELTA scores (age 73.7 ± 7.1 years, 55.4% female, 100% white/Caucasian). Higher baseline DELTA scores (stronger evidence of cognitive decline) predicted higher baseline CDR-SOB ($\Delta R^2 = .318$) and faster rates of CDR-SOB increase over time ($\Delta R^2 = .209$). Longitudinal changes in DELTA scores tracked closely and in the same direction as CDR-SOB scores (fixed and random effects of mean + mean-centered DELTA, $\Delta R^2 > .7$). Results were similar for FAQ scores. High DELTA scores predicted higher PET-A β SUVR ($\rho = .324$), higher CSF-pTau/CSF-A β ratio ($\rho = .460$), and demonstrated PPV > .9 for positive Alzheimer's disease biomarker classification. **Conclusions:** Data support initial development and validation of the DELTA score through its associations with longitudinal functional changes and Alzheimer's biomarkers. We provide several considerations for future research and include an automated scoring program for clinical use.

Keywords: DELTA, Assessment, Cognitive decline, Psychometrics, ADNI, Base rate

INTRODUCTION

Neuropsychological assessments are the accepted standard-of-care for measuring cognition. Many of the most common

neuropsychological tests have existed for decades but research on their strengths and limitations has led to improvements in how they are used and interpreted in clinical, research, and forensic settings. For example, research

*Correspondence and reprint requests to: Breton M. Asken, Alpert Medical School of Brown University, Department of Psychiatry and Human Behavior, University of Florida, Department of Clinical and Health Psychology, Box G-BH, Division of Clinical Psychology, Providence, RI 02912, USA. Tel: +1 401 444 1929; Fax: +1 401 444 1911. E-mail: basken8@php.ufl.edu

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

This is an updated version of the original article. For details please see the notice at <https://doi.org/10.1017/S1355617720000557>.

showing the base rates with which cognitively intact individuals achieve low scores across a test battery has helped reduce false positive conclusions that a patient has declined (Binder, Iverson, & Brooks, 2009; Brooks & Iverson, 2010; Brooks, Iverson, & White, 2009). There is also evolving awareness of the complex relationships between diverse neuropathologic changes and heterogeneous cognitive phenotypes (Boyle et al., 2018; James et al., 2016; Wennberg et al., 2019).

Pioneering work from Bondi and Jak in longitudinal aging cohorts rather consistently demonstrates that actuarial approaches that classify cognitive impairment using patterns and frequencies of low scores have led to modest rates of clinical reversion (mild cognitive impairment or “MCI” to “cognitively normal” at follow-up), improved characterization of risk of progression to dementia, and stronger associations with biologic disease markers than “single-test” methods (Bondi et al., 2014; Bondi & Smith, 2014; Jak et al., 2009; Jak et al., 2016; Petersen et al., 1999). Oltra-Cucarella et al. (2018) recently showed that the number of low scores in a test battery predicted progression from MCI to dementia in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort with improved specificity (Oltra-Cucarella et al., 2018). Put simply, implementing these approaches increases a clinician’s confidence about whether a patient’s test scores reflect true cognitive changes *versus* normal performance variability unrelated to suspected underlying disease.

Deemphasizing individual test scores known to fluctuate in cognitively normal individuals promotes clinical translation by more closely mimicking the holistic interpretations used by neuropsychologists. One possible limitation of these methods, however, is that they dichotomize impairment status rather than allowing for a continuum of evidence for cognitive decline. Dichotomous approaches that reduce complex cognitive profiles derived from psychometrically imperfect measures might not classify patients into discrete diagnostic groups accurately or reliably. Moving away from dichotomous categorizations (“not impaired” vs. “impaired”) and toward a continuous spectrum construct might advance clinical assessment methods further and promote integration with similarly complex disease biomarker measures.

Data from large longitudinal research cohorts with biomarker collection challenge traditional conceptualizations of disease–phenotype relationships and underscore the imperfect alignment of disease states and clinical syndromes (Jack et al., 2018). For example, biomarker [e.g., positron emission tomography (PET)] evidence for amyloid plaque and tau tangle Alzheimer’s disease (AD) pathology does not guarantee measurable cognitive impairment (De Meyer et al., 2010; Mortamais et al., 2017) and, when present, cognitive impairment is not universally the prototypic AD presentation of “rapid forgetting” (Ossenkoppele et al., 2015; Perry et al., 2017; Phillips et al., 2018). Yet, in the absence of advanced biomarkers of disease pathology, clinical presentation alone may not differentiate disease states adequately (e.g., amnesic profiles associated with both AD

and limbic-predominant age-related TDP-43 encephalopathy) (Nelson et al., 2019). The National Institute on Aging and Alzheimer’s Association (NIA-AA) established the “A/T/N” framework for biomarker evidence of AD with expected adaptation to include additional neuropathologic biomarkers (Jack et al., 2018; Nelson et al., 2019). Anticipating this paradigm shift, a cognitive correlate derived from neuropsychological evaluations would complement biomarker frameworks by systematically quantifying evidence of domain-specific cognitive decline.

The purpose of this study is to develop and validate the Discrepancy-based Evidence for Loss of Thinking Abilities (DELTA) score. In the absence of prior test scores for comparison, DELTA scores characterize evidence for cognitive decline on a continuous spectrum based on the extent of discrepancies between obtained test scores and predicted pre-morbid scores derived from multiple-variable regression models. The approach reflects the progress of prior research demonstrating the benefits of accounting for low-score-base-rates among cognitively normal individuals and psychometric principles for improving detection of cognitive changes (Iverson & Brooks, 2011). This initial validation used the ADNI cohort and evaluated how the DELTA score predicted functional changes over time, as well as its association with AD biomarkers. We provide a Microsoft Excel-based scoring program that directly incorporates the study findings and we discuss next steps for broader validation outside of AD samples.

METHODS

Data Source and Participants

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. ADNI was approved by the institutional review boards of all participating institutions. Informed written consent was obtained from all participants at each site. Data included in our study were participant demographics, neuropsychological test results, functional measures including the Clinical Dementia Rating scale (CDR) and Functional Activities Questionnaire (FAQ), and biomarkers of beta-amyloid[A β ; via PET and cerebrospinal fluid (CSF)] and phosphorylated tau (p-tau; via CSF).

The CDR (Morris, 1993) has both a global score (range 0–3, where 0 = “normal” and 3 = “severe dementia”) and Sum of Boxes (SOB) score (range 0–18) that quantify aspects of daily functioning including memory, orientation, judgment/problem-solving, community affairs, home and hobbies, and personal care. The FAQ (Pfeffer, Kurosaki,

Table 1. Regression equations for predicting test scores based on age, gender, years of education, and word-reading ability (ANART total errors). Scores within the table used for DELTA score calculation include AVLT Delayed Recall, LM Delayed Recall, Trails B/Trails A, Animal Fluency Total Correct, and BNT-30 Total Correct. Regression equations not calculated for Clock Drawing due to limited score range in control group (4 or 5). The “ANART R^2 ” column indicates the added variance attributed specifically to word-reading performance above and beyond age, gender, and years of education

Test	Test component	Predicted raw score equation	SEE	Model R^2	ANART R^2
AVLT	Total Learning ^{a,b}	$77.107 + (\text{Age}^*-.464) + (\text{Gender}^*-.5.10) + (\text{YrsEduc}^*.599) + (\text{ANART errors}^*.068)$	9.12	.15	.001
	Trial 6 ^{a,b}	$16.91 + (\text{Age}^*-.138) + (\text{Gender}^*-.1.51) + (\text{YrsEduc}^*.230) + (\text{ANART errors}^*.054)$	3.42	.10	.007
	Delayed Recall ^{a-c}	$14.53 + (\text{Age}^*-.133) + (\text{Gender}^*-.1.31) + (\text{YrsEduc}^*.284) + (\text{ANART errors}^*.060)$	3.70	.08	.008
LM	Immediate Recall ^{c,d}	$12.50 + (\text{Age}^*-.001) + (\text{Gender}^*-.72) + (\text{YrsEduc}^*.233) + (\text{ANART errors}^*-.090)$	2.99	.09	.03
	Delayed Recall ^{c,d}	$11.33 + (\text{Age}^*.001) + (\text{Gender}^*-.80) + (\text{YrsEduc}^*.249) + (\text{ANART errors}^*-.089)$	3.15	.09	.02
Trail Making Test	Trails A ^a	$3.72 + (\text{Age}^*.483) + (\text{Gender}^*-.1.72) + (\text{YrsEduc}^*-.193) + (\text{ANART errors}^*-.078)$	10.83	.06	.002
	Trails B ^a	$-76.39 + (\text{Age}^*1.97) + (\text{Gender}^*-.4.13) + (\text{YrsEduc}^*.41) + (\text{ANART errors}^*.814)$	32.17	.12	.02
	Trails B/Trails A ^d	$.619 + (\text{Age}^*.017) + (\text{Gender}^*-.078) + (\text{YrsEduc}^*.025) + (\text{ANART errors}^*.032)$.938	.05	.03
Clock Drawing	–	–	–	–	–
Animal Fluency	Total Correct ^c	$26.67 + (\text{Age}^*-.146) + (\text{Gender}^*-.394) + (\text{YrsEduc}^*.388) + (\text{ANART errors}^*-.063)$	5.184	.07	.006
BNT-30	Total Correct ^{b,d}	$31.12 + (\text{Age}^*-.048) + (\text{Gender}^*1.362) + (\text{YrsEduc}^*-.050) + (\text{ANART errors}^*-.063)$	1.946	.13	.02

ANART, American National Adult Reading Test; AVLT, Rey Auditory Verbal Learning Test; BNT, Boston Naming Test; LM, Logical Memory (WMS-R); SEE, standard error of the estimate; YrsEduc, Years of Education.

^a Age $p < .01$.

^b Gender $p < .01$.

^c Years of Education $p < .01$.

^d ANART Error Score $p < .01$.

Harrah Jr., Chance, & Filios, 1982) assesses the degree of assistance individuals need when completing 10 instrumental activities of daily living (range 0–30 with higher scores representing greater assistance needed). Lastly, A β and p-tau are the hallmark neuropathologic features of AD under the biologic A/T/N classification system and commonly underlie cognitive and behavioral changes in older adults (Jack et al., 2018).

The ADNI cohort was used as the normative reference for DELTA development. Regression coefficients were derived from a sample of robust cognitively normal (RCN) individuals. RCNs had to have CDR = 0 and Mini Mental Status Exam (MMSE) score ≥ 29 at both baseline and 1-year follow-up. We regressed age, gender, years of education, and word-reading ability against each of the neuropsychological test scores. Equations derived from the results of these regression models were then applied to the entire ADNI cohort to compute individual participants' predicted premorbid test scores. The discrepancy between a participants' predicted and obtained scores is the key component of the DELTA score, which theoretically represents the likelihood of true cognitive decline from a predicted baseline state.

Discrepancy-Based Evidence for Loss of Thinking Abilities (DELTA) Score Development

The DELTA score is broadly based on the following: (1) the degree of discrepancy between an individual's predicted and observed scores on individual tests within a battery, and (2) the frequency of discrepancy scores that exceed common cut-offs for infrequently occurring or “impaired” scores. This initial development and validation is specific to ADNI test scores and represents a proof of concept.

Step 1: Identifying ADNI test scores to incorporate into DELTA

The six tests used for the ADNI-based DELTA spanned three cognitive domains: Memory [Rey Auditory Verbal Learning Test (AVLT), Wechsler Memory Scale – Revised Logical Memory (WMS-R LM)], Language [Animal Fluency, 30-item Boston Naming Test (BNT-30)], and Executive Function (Clock Drawing, Trails B); see Supplemental Table A for test details. We isolated the “executive” component of Trails B by dividing Trails B time by Trails A time to reduce confounding

effects of psychomotor speed on measuring the set-shifting component of Trails B (Arbuthnott & Frank, 2000).

Step 2: Calculating test score prediction equations from robust cognitively normal (RCN) participants

Predicted scores for each relevant test component came from regression equations using coefficients (B-weights) corresponding to the effects of age, gender (male or female), years of education, and word-reading ability on test scores in the RCN subgroup (Eppig et al., 2017). Word-reading scores came from the American National Adult Reading Test (ANART number of errors), which estimates general intelligence (i.e., IQ) and informs expected premorbid cognitive abilities (McGurn et al., 2004). Word-reading ability was chosen as a performance-based predictor of cognitive abilities to improve upon typical demographic-only adjustment methods (Crawford, Moore, & Cameron, 1992; Duff, Chelune, & Dennett, 2011; Duff, Dalley, Suhrie, & Hammers, 2018). All predictors were left in the equation regardless of statistical significance; this approach captures any variance explained by these commonly collected variables and more clearly allows for direct comparison of their relative prediction strengths across test scores. The test score-specific prediction equations therefore looked like:

$$\begin{aligned} \text{Predicted Score} = & \text{Constant} + (\text{Age} * B_{\text{Age}}) \\ & + [\text{Gender}(1 \text{ or } 2) * B_{\text{Gender}}] \\ & + (\text{Years of Education} * B_{\text{YrsEducation}}) \\ & + (\text{ANARTerrors} * B_{\text{ANARTerrors}}) \end{aligned}$$

Step 3: Calculating standardized discrepancy scores

We identified 270 ADNI participants as RCNs. We first calculated predicted raw scores for RCNs and then standardized by z-transforming the discrepancy (z-Discrep) between predicted and actual raw scores by dividing the difference by the test-specific regression model's standard error of the estimate (SEE; defined as the standard deviation (SD) of the error term). This occurred for all test scores (Table 1) except for Clock Drawing because all RCNs obtained scores of either 4 or 5 (out of 5). We subtracted actual from predicted scores for the Trail Making Test so that negative z-scores reflected poor performance.

$$z - \text{Discrep} : \text{Standardized}(z)\text{Discrepancy} = \frac{(\text{Actual Test Score} - \text{Predicted Test Score})}{\text{SEE}}$$

Step 4: Defining cutoffs for evidence of cognitive decline

A key component to neuropsychological test score interpretations is the frequency with which a given score occurs in a reference population (e.g., scores corresponding to a z-score of -2.0 are atypically low and usually interpreted as strong evidence of cognitive decline). We therefore established

percentile cutoffs for infrequently occurring z-Discrep scores: 16th, 7th, and 2nd percentile. We used percentiles due to non-normality of the z-Discrep score distributions. These percentiles correspond to commonly used cutoffs in normally distributed data (16th%ile for $z = -1.0$, 7th%ile for $z = -1.5$, 2nd%ile for $z = -2.0$) (Iverson & Brooks, 2011).

Step 5: Defining DELTA score criteria

We based DELTA score criteria on the principle that obtaining low scores on Test A and Test B within a cognitive domain occurs less frequently than obtaining low scores on Test A or Test B (Table 2). The DELTA score also accounts for the degree of discrepancy between obtained and predicted scores. For example, if both the BNT and Animal Fluency scores have a z-Discrep below the second percentile, the individual receives a Language score of 5. However, if only one of those two z-Discrep scores is below the second percentile and the other is normal, they receive a Language score of 3. The same score of 3 could also be obtained if both z-Discrep scores fall between the second and seventh percentile.

A unique component of the Memory DELTA score for both AVLT and LM is the requirement of <50% retention of initially learned information. This was done to reduce confounding effects of poor immediate recall on delayed recall scores due to non-memory factors like inattentiveness or executive deficits (Casaletto et al., 2017). We chose <50% retention arbitrarily and this cutoff was applied uniformly.

Step 6: Calculating the DELTA score

Domain-specific scores range from 0 to 5. A maximum domain score of 5 corresponds to both z-Discrep scores within that domain falling below the second percentile (note that the cutoff value of which depends on whether the individual had a "Low," "Average," or "High" predicted score - see Results). For the Memory domain, this also requires <50% retention on both AVLT and LM delayed recall. Domain-specific scores are then summed for a total DELTA score (0–15).

Methods and Analyses for Validating DELTA Scores

We calculated DELTA scores for all ADNI participants with complete test data at Time 0 (baseline assessment; BL) and for each follow-up year up to the fifth year (Y1, Y2, Y3, Y4, Y5). Functional outcomes included the CDR-SOB and FAQ scores corresponding to time points with calculable DELTA scores. Functional outcomes were used as the primary validation of the DELTA score because it is a purely clinical and psychometrically based score that is independent of biomarker indicators of specific diseases processes. Biomarkers were used in the secondary validation.

Table 2. Criteria for calculating the DELTA score (automatic scoring program provided)

Domain	Code	Criterion	Risk points	Max possible
Memory	M.1	Both AVLT and LM Delayed Recall z-Discrep < 2nd%ile + %Savings < 50%	5	5
	M.2	a) Either AVLT or LM Delayed Recall z-Discrep < 2nd%ile + %Savings < 50% b) Both AVLT and LM Delayed Recall 2nd%ile < z-Discrep < 7th%ile	3	
	M.3	Either AVLT or LM Delayed Recall 2nd%ile < z-Discrep < 7th%ile + %Savings < 50%	2	
	M.4	Both M.2a and M.3 criteria met (only possible for one test to meet one criterion)	4	
	M.5	Both AVLT and LM Delayed Recall 7th%ile < z-Discrep < 16th%ile + %Savings < 50%	1	
	M.6	All other score combinations	0	
Language	L.1	Both BNT and Animal Fluency z-Discrep < 2nd%ile	5	5
	L.2	a) Either BNT or Animal Fluency z-Discrep < 2nd%ile b) Both BNT and Animal Fluency 2nd%ile < z-Discrep < 7th%ile	3	
	L.3	Either BNT or Animal Fluency 2nd%ile < z-Discrep < 7th%ile	2	
	L.4	Both L.2a and L.3 criteria met (only possible for one test to meet one criterion)	4	
	L.5	Both BNT and Animal Fluency 7th%ile < z-Discrep < 16th%ile	1	
	L.6	All other score combinations	0	
Executive	E.1	Both Clock Drawing = 0–2 and (Trails B time/Trails A time) z-Discrep < 2nd%ile	5	5
	E.2	a) Either Clock Drawing = 0–2 or (Trails B time/Trails A time) z-Discrep < 2nd%ile b) Both Clock Drawing = 3 and (Trails B time/Trails A time) 2nd%ile < z-Discrep < 7th%ile	3	
	E.3	Either Clock Drawing = 3 or (Trails B time/Trails A time) 2nd%ile < z-Discrep < 7th%ile	2	
	E.4	Both E.2a and E.3 criteria met (only possible for one test to meet one criterion)	4	
	E.5	Both Clock Drawing = 4 and (Trails B time/Trails A time) 7th%ile < z-Discrep < 16th%ile	1	
	E.6	All other score combinations	0	
Total Risk Score			X/15	15

Primary Validation with Functional Outcomes

First, we examined DELTA score changes in the RCN group and the entire baseline sample at follow-up.

Second, we used linear mixed model analyses with maximum likelihood estimation evaluating associations between DELTA score and longitudinal functional changes. Model fit was evaluated in a hierarchical (i.e., nested) approach relative to the unconditional means (null) model: (1) fixed and random effect of time, (2) fixed effects of age, gender, and years of education, (3) fixed effect of BL DELTA score, (4) fixed and random effect of BL DELTA \times Time interaction. This first approach most closely mimics a clinical scenario where a patient obtains a DELTA score and the clinician wants to know how that predicts future everyday functioning.

Third, we leveraged the longitudinal cognitive data in ADNI by building the following model using DELTA as a time-varying covariate: (1) fixed and random effect of time, (2) fixed effects of age, gender, and years of education, (3) fixed effects of mean-DELTA and mean-centered-DELTA (decoupled to control for each case's mean cognitive functioning across the study), (4) random effect of mean-centered DELTA. This second approach examines how well changes in DELTA scores coincide with changes in functional outcomes over time.

Separate analyses were run with CDR-SOB and FAQ scores as the dependent variable. We tracked overall model fit using $-2 \text{ Log Likelihood}$ and Akaike's Information Criterion changes at each step as well as reductions in unexplained variance (R^2 change) for the covariance parameters (random effects).

Secondary Validation with Alzheimer's Disease Biomarkers

We examined associations between DELTA scores, PET-A β burden, and CSF evidence of AD, stratifying participants by apolipoprotein E (APOE) e4 carriers and noncarriers. PET-A β was quantified using PET scanning with 18F-florbetapir (AV45) tracer. Standardized uptake value ratios (SUVr) were calculated by ADNI by dividing mean cortical florbetapir uptake (frontal, anterior/posterior cingulate, lateral parietal, lateral temporal) by whole cerebellar uptake. PET-A β positivity reflected a cross-sectional SUVr > 1.11 (Landau et al., 2014). CSF evidence of AD was determined by cutoff scores optimized for ADNI (Hansson et al., 2018) that used the CSF-hyperphosphorylated tau (CSF-pTau) to CSF-A β (1–42) ratio (CSF-pTau/CSF-A β > .0251 pg/ml). We analyzed continuous associations between DELTA score and biomarker burden using Spearman's rho for non-normal data and examined

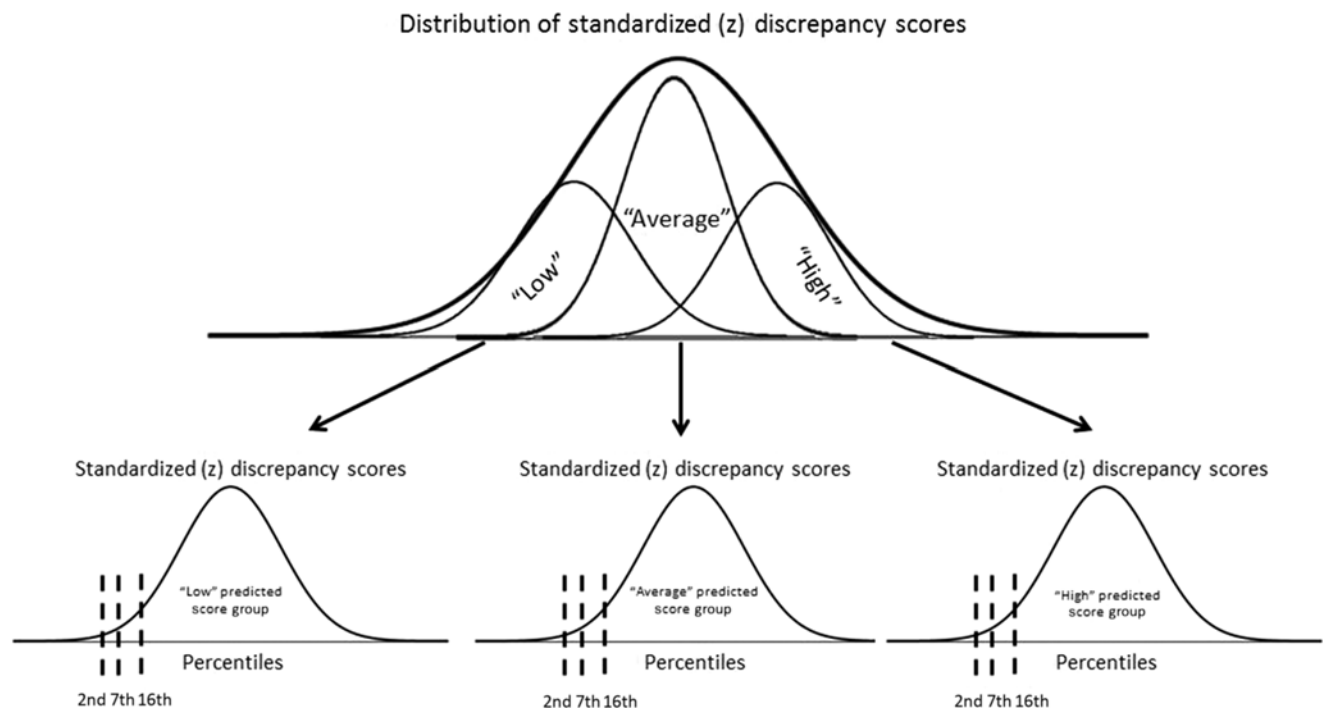


Fig. 1. Conceptual figure demonstrating derivation of the z-Discrep score that corresponds to a given percentile cutoff, stratified by each participant's predicted raw test score ("Low," "Average," or "High"). All bell curves represent the theoretical distribution of standardized (z) discrepancy scores. The top curve shows the z-Discrep distribution for the entire RCN sample along with the theoretical "Low," "Average," and "High" predicted score subgroups that make up the overall RCN sample. The bottom curves show that these subgroups were isolated so that the z-Discrep scores that correspond to the 2nd, 7th, and 16th percentile cutoffs would be specific to the predicted score group's z-Discrep distribution.

the positive predictive value (PPV) of a given DELTA group for dichotomized biomarker outcomes (PET-A β positivity and CSF-AD positivity). See www.loni.usc.edu for acquisition and processing details.

All statistical analyses were performed using SPSS v.22 or v.25. A priori alpha levels were set at $p < .005$ (unless otherwise noted) to partially account for spurious findings associated with a large sample size and to reflect recent proposals to lower thresholds for enhancing replication of new discoveries (Benjamin et al., 2018; Ioannidis, 2018).

Restricted Sample

The ADNI sample considered for this study was around 92% white/Caucasian (4% black/African American, 2% Asian, <1% each of multiple races, American Indian/Alaskan, and Hawaiian). We elected to restrict DELTA score development and validation to white/Caucasian participants and openly acknowledge this limited scope. We assumed that indiscriminately applying the data underlying DELTA score development across diverse racial and ethnic groups was inappropriate. The well-documented and complex relationships between sociodemographic factors and cognitive test scores (Dotson, Kitner-Triolo, Evans, & Zonderman, 2009; Manly & Echemendia, 2007; Rivera Mindt, Byrd, Saez, & Manly, 2010) require careful consideration when

extrapolating data from unrepresentative samples (Brooks, Sherman, Iverson, Slick, & Strauss, 2011). We hope these methods will be replicated using racially and ethnically diverse cohorts.

RESULTS

The RCN group included 270 participants (mean \pm SD age = 74.7 ± 5.5 years, 51.1% female, 100% white/Caucasian, 74.1% APOE ϵ 4 noncarriers; MMSE mean \pm SD = $29.6 \pm .5$, education years \pm SD = 16.7 ± 2.6). Regression-based equations for predicting test score performance (including some not used for DELTA score calculation), as well as the added variance explained by the word-reading ability component, are provided as Table 1.

The z-Discrep score distributions varied in the RCN group as a function of predicted score. Those with higher predicted test scores had a different distribution of z-Discrep scores than those with lower predicted test scores. For example, z-Discrep = -1.86 corresponds to the seventh percentile for participants with high predicted AVLT delayed recall, whereas z-Discrep = -1.42 corresponds to the seventh percentile for those with low predicted AVLT delayed recall. Therefore, we stratified the predicted scores for each test into "Low" (lower quartile), "Average" (middle quartiles), and "High" (upper quartile) groups and then identified the z-

Discrep values corresponding to the 16th, 7th, and 2nd percentile that were specific to the predicted score group (Figure 1).

We created five “Level of Evidence” for cognitive impairment groups based on the DELTA score distribution in the RCN sample: DELTA = 0 (“No Evidence” of cognitive decline; 73.8% of RCNs), DELTA = 1–3 (“Low Evidence”; 24.6% of RCNs), DELTA = 4–6 (“Moderate Evidence”; 1.6% of RCNs), DELTA = 7–9 (“Strong Evidence”; .0% of RCNs), DELTA = 10+ (“Very Strong Evidence”; .0% of RCNs).

Evidence for Incremental Value of DELTA Score Use

Addition of word-reading ability as a performance-based predictor of premorbid performance significantly improved model fit in 3 out of the 5 predicted test scores derived from regression equations (Clock Drawing excluded) for DELTA score calculation: LM Delayed Recall, Trails B/A proportion score, and the BNT-30 (Table 1). Word-reading independently accounted for 15–60% of the total variance explained by the overall models for these tests.

We then examined the potential benefit of using multiple test scores for characterizing cognitive abilities within a domain by comparing rates of “low scores” on single tests to frequencies of DELTA scores. Individual test scores were converted to percentiles based on the RCN score distribution. Using memory tests and a seventh percentile cutoff ($z < -1.5$ in a normal distribution) as an exemplar, we observed that 10.5% of RCNs had a LM Delay score <7th%ile and 9.7% had an AVLT Delay score <7th%ile, while 19.1% had *either* one or the other. In other words, one in five cognitively intact individuals may get flagged as having “impaired” memory if relying on individual test scores. However, 97% of the RCN sample had a DELTA Memory score of 0 and 98% had an overall DELTA score in the “No Evidence” (DELTA = 0) or “Low Evidence” range (DELTA = 1–3). This suggests a possible reduction in “false positive” determinations of cognitive decline using the DELTA score.

Longitudinal DELTA Scores

Inspection of the overall sample’s longitudinal mean functional and DELTA score data showed expected group-level worsening until around Y3 and then decreased scores between Y3 and Y4 that held through Y5, suggesting a survival bias in the sample’s attrition over time. There was also a consistent decline in representation of APOEε4 carriers. We therefore focused results on longitudinal data spanning BL, Y1, Y2, and Y3. Table 3 shows descriptive data stratified by assessment time point.

Of the 270 RCNs used for calculating regression-predicted test score performance, 256 (94.8%) completed all tests and therefore had calculable DELTA scores. Rates of DELTA group changes for the RCN sample and overall BL sample are provided as supplemental material

(Supplemental Table B) and in Table 4. Over 83% of RCNs with “No Evidence” at baseline *and follow-up cognitive data* (i.e., *remained in the study*) stayed in this group at Y1, Y2, and Y3 follow-up. For RCNs with “Low Evidence” at BL and with follow-up cognitive data, over 90% remained either as “Low Evidence” or reverted to “No Evidence.”

For the overall BL sample (Table 4), progression to higher levels of evidence for cognitive decline (e.g., “Moderate” or higher) was relatively rare for those with “No Evidence” at BL, but rates increased as a function of BL DELTA group. Reversion and progression percentages become skewed at the highest BL DELTA groups (“Strong” and “Very Strong”) due to lower rates of these levels of evidence at BL and greater loss to follow-up.

DELTA Validation Against Longitudinal Functional Changes

Linear mixed model analyses focused on CDR-SOB and FAQ changes from BL through Y3 as a function of BL DELTA scores. Table 5 shows the model fit characteristics at each stage of the analyses. Higher BL DELTA score (i.e., worse cognition) predicted higher BL CDR-SOB beyond the effects of age, gender, and education (between-case intercept, $\Delta R^2 = .318$, large effect). The BL DELTA score \times Time interaction term significantly improved model fit ($\Delta R^2 = .742$, large effect) and suggested that higher BL DELTA score was associated with faster increases in CDR-SOB score (i.e., worsening) over time. No covariates explained additional within-case residual variance (i.e., deviation from regression-predicted CDR-SOB) above the effects of time ($\Delta R^2 < .02$).

We then ran the models using participants’ longitudinal, visit-specific DELTA scores instead of just the BL DELTA score (i.e., DELTA as a time-varying covariate). Higher mean and mean-centered BL DELTA scores predicted faster rates of increase in CDR-SOB score beyond the effects of time (occasion-intercept, $\Delta R^2 = .658$, large effect), and accounting for person-specific longitudinal changes in DELTA score further improved the model ($\Delta R^2 = .791$). These results suggest that longitudinal changes in CDR-SOB track closely and in the same direction as changes in DELTA score. Of note, remaining significant unexplained variance indicated that the degree to which DELTA score tracks with CDR-SOB is not uniform across all participants (i.e., the strength of the association between DELTA score and CDR-SOB differs from person to person).

Results were similar when evaluating longitudinal FAQ changes in place of CDR-SOB. For all analyses, age, gender, and years of education did not significantly predict longitudinal functional changes in models that included DELTA scores.

DELTA Validation Against PET and CSF Biomarkers

Biomarker validation was performed on the subset of the BL sample with available PET and/or CSF biomarkers.

Table 3. Sample descriptive statistics stratified by assessment point. Robust cognitively normal (RCN) data represent the RCN samples baseline visit

	RCN	Baseline	Year 1	Year 2	Year 3	Year 4	Year 5
<i>n</i>	256	1359	1070	915	576	474	234
Age, y							
Mean (SD)	74.7 (5.5)	73.7 (7.1)	74.6 (7.0)	75.4 (7.0)	76.0 (7.1)	76.8 (7.0)	77.9 (6.9)
Median (IQR)	73.9 (70.9–78.2)	73.6 (69.2–78.8)	74.7 (70.2–79.5)	75.5 (70.8–80.2)	76.1 (71.2–80.9)	76.7 (72.0–81.6)	78.0 (74.9–82.7)
Gender (%female)	50.4	44.6	43.3	45.5	43.2	43.0	39.7
Education, y							
Mean (SD)	16.7 (2.6)	16.0 (2.8)	16.1 (2.8)	16.2 (2.7)	16.2 (2.7)	16.3 (2.6)	16.7 (2.8)
Median (IQR)	16 (15–19)	16 (14–18)	16 (14–18)	16 (14–18)	16 (14–18)	16 (14–18)	16 (14–18)
APOEε4 (% Carriers)	26.1	44.7	43.7	40.4	37.9	36.9	32.1
CDR-SOB							
Mean (SD)	0 (.1)	1.4 (1.6)	1.6 (1.9)	1.4 (2.0)	1.5 (2.1)	1.3 (2.0)	1.2 (1.6)
Median (IQR)	0 (0–0)	1.0 (0–2.0)	1.0 (0–2.5)	.5 (0–2.0)	1.0 (0–2.0)	.5 (0–2.0)	.5 (0–2.0)
Min.–Max.	0–1.0	0–10.0	0–12.0	0–16.0	0–14.0	0–17.0	0–7.0
FAQ Total							
Mean (SD)	.2 (.7)	3.4 (5.3)	4.2 (6.4)	3.6 (6.2)	3.8 (6.4)	3.2 (5.8)	3.0 (5.4)
Median (IQR)	0 (0–0)	1 (0–5)	1 (0–6)	0 (0–4)	0 (0–5)	0 (0–3)	0 (0–4)
Min.–Max.	0–6	0–28	0–30	0–30	0–29	0–29	0–27
Neuropsychological Testing [Median (IQR)]							
LM I	14 (12–17)	10 (6–13)	11 (7–15)	12 (8–16)	12 (8–15)	13 (10–16)	14 (10–16)
LM II	13 (11–16)	8 (3–11)	9 (3–13)	10 (5–15)	10 (5–14)	12 (7–15)	12 (7–15)
LM % Retention	93 (85–100)	79 (50–94)	80 (50–94)	85 (64–100)	85 (63–100)	88 (70–100)	90 (73–100)
AVLT (Trial “6”)	9 (6–11)	5 (3–9)	6 (3–9)	7 (3–10)	6 (3–9)	7 (4–10)	7 (4–10)
AVLT Delay	8 (5–11)	4 (1–8)	4 (0–8)	5 (1–9)	5 (1–8)	5 (2–10)	5 (1–8)
AVLT % Retention	91 (75–100)	78 (33–100)	75 (0–100)	80 (25–100)	80 (41–100)	83 (40–100)	78 (40–100)
Clock Drawing	5 (5–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)
Trails B/A Ratio	2.2 (1.8–2.8)	2.5 (2.0–3.4)	2.5 (2.0–3.3)	2.5 (2.0–3.3)	2.5 (2.0–3.2)	2.4 (2.0–3.2)	2.5 (2.0–3.2)
Animal Fluency	21 (17–24)	18 (14–22)	18 (14–22)	19 (14–23)	18 (14–21)	19 (15–22)	18 (15–22)
BNT-30	29 (28–30)	28 (25–29)	29 (26–30)	29 (27–30)	29 (26–30)	29 (27–30)	29 (27–30)

APOEε4, apolipoprotein epsilon 4; AVLT, Rey Auditory Verbal Learning Test; BNT-30, 30-item Boston Naming Test; CDR-SOB, Clinical Dementia Rating-Sum of Boxes score; FAQ, Functional Activities Questionnaire; IQR, interquartile range; LM, Logical Memory (WMS-R); Min.–Max., minimum value–maximum value; RCN, robust cognitively normal; SD, standard deviation; y, years.

Table 4. Change in DELTA group status based on BL DELTA group for the entire BL sample. Values represent the percentage of participants with follow-up DELTA scores within each DELTA group. Interpret reversion and progression percentages at the highest BL DELTA groups (“Strong” and “Very Strong”) with caution due to lower frequency of these DELTA groups at BL and greater loss to follow-up (i.e., survivor bias), particularly by Year 3

Overall baseline DELTA group (score range)	Follow-up DELTA group (score range)	Year 1	Year 2	Year 3
“Level of Evidence”	“Level of Evidence”	<i>n</i> = 497	<i>n</i> = 490	<i>n</i> = 300
No Evidence (0) <i>n</i> = 593	No Evidence (0)	77.1	74.1	76.3
	Low (1–3)	19.1	20.6	19.0
	Moderate (4–6)	3.2	4.5	3.7
	Strong (7–9)	.6	.6	.7
	Very Strong (10–15)	.0	.2	.3
Low (1–3) <i>n</i> = 379		<i>n</i> = 303	<i>n</i> = 270	<i>n</i> = 190
	No Evidence (0)	44.2	41.1	39.5
	Low (1–3)	33.0	34.4	33.7
	Moderate (4–6)	16.8	17.0	16.3
	Strong (7–9)	5.0	5.9	8.4
Moderate (4–6) <i>n</i> = 249		<i>n</i> = 169	<i>n</i> = 109	<i>n</i> = 78
	No Evidence (0)	9.5	14.7	16.7
	Low (1–3)	23.7	31.2	17.9
	Moderate (4–6)	34.9	26.6	37.2
	Strong (7–9)	27.2	15.9	20.5
Strong (7–9) <i>n</i> = 141		<i>n</i> = 78	<i>n</i> = 49	<i>n</i> = 18
	No Evidence (0)	2.6	2.0	5.6
	Low (1–3)	5.1	8.2	11.1
	Moderate (4–6)	23.1	14.3	33.3
	Strong (7–9)	38.5	32.7	22.2
Very Strong (10–15) <i>n</i> = 57		<i>n</i> = 26	<i>n</i> = 12	<i>n</i> = 3
	No Evidence (0)	.0	.0	.0
	Low (1–3)	3.8	16.7	.0
	Moderate (4–6)	7.7	8.5	33.3
	Strong (7–9)	26.9	25.0	33.3
	Very Strong (10–15)	61.5	50.0	33.3

DELTA, Discrepancy-based Evidence for Loss of Thinking Abilities.

Higher BL DELTA score predicted higher PET-A β SUVR (*n* = 739, ρ = .324, medium effect), lower CSF-A β (*n* = 1000, ρ = -.412, medium–large effect), higher CSF-pTau (*n* = 998, ρ = .340, medium effect), and higher CSF-pTau/CSF-A β ratio (*n* = 997, ρ = .460, medium–large effect); all *p*'s < .001. We looked at the Memory subscore (0–5) of the DELTA score independently and found similar relationships with biomarkers as the total DELTA score.

Figure 2A shows relationships among DELTA groups and PET-A β status stratified by APOE ϵ 4 noncarriers and carriers. Among APOE ϵ 4 noncarriers with PET-A β scans (*n* = 416), 135 (32.5%) were PET-A β (+). Positive predictive value (PPV) of the DELTA scores increased as a function of DELTA “level of evidence” group from 25.7% in the DELTA = 0 group (“No Evidence”) to 63.6% for participants with DELTA > 6 (“Strong Evidence” plus “Very Strong Evidence” groups). Results were similar when looking at

Memory score only (Figure 2B). Among APOE ϵ 4 carriers with PET-A β scans (*n* = 320), 240 (75.0%) were PET-A β (+). PPV of the DELTA groups increased from 62.7% in the DELTA = 0 group (“No Evidence”) to 92.7% for participants with DELTA > 3 (“Moderate Evidence” or higher groups). We found slightly stronger relationships based on the Memory subscore, such that a Memory scores of 4 (17/17 participants) and 5 (17/17 participants) had 100% PPV.

Figure 3A shows relationships among DELTA groups and CSF-AD biomarker status stratified by APOE ϵ 4 noncarriers and carriers. Among APOE ϵ 4 noncarriers with CSF-AD biomarkers (pTau/A β ratios, *n* = 553), 166 (30.0%) were CSF-AD(+). PPV of the DELTA scores again increased as a function of the DELTA “level of evidence” group from 18.6% in the “No Evidence” group to 88.9% in the “Very Strong Evidence” group (8/9 participants). Results were similar when looking at Memory score only (Figure 3B). Among

Table 5. Linear mixed model analysis of baseline and longitudinal DELTA scores predicting CDR-SOB and FAQ score changes over 3-year follow-up. Terms: “Between-Intercept” – Variance associated with between-participant baseline differences (i.e., initial CDR-SOB/FAQ score); “Within-Residual” – Variance associated with discrepancies between regression-predicted and actual CDR-SOB/FAQ score for each participant; “Time-Intercept” – Variance associated with rates of change in CDR-SOB/FAQ score over time (i.e., between-participant differences in slope of change)

Linear mixed model tracking	Model parameters			Model fit statistics			Unexplained variances and overall model R^2					
	Model predictors	Fixed or random	Parameters (n)	-2LL	ΔX^{2a}	AIC	Between-intercept	R^2	Within-residual	R^2	Time-intercept	R^2
CDR-SOB Model												
CDR-SOB Base Model	Unconditional Means	–	3	19376	–	19382	4.56	–	1.92	–	–	–
	+Time	Fixed + Random	5	17324	2051.4	17334	2.93	.357	.585	.695	.971	–
	+Age/Gender/Education	Fixed	8	17279	45.1	17295	2.84	.378	.585	.695	.969	.002
CDR-SOB Base Model + Baseline DELTA Score Covariates												
Baseline DELTA Effects	+BL DELTA Score	Fixed	9	15448	1831.7	15466	1.39	.696	.581	.697	.768	.209
	+BL DELTA \times Time	Fixed + Random	11	14669	779.4	14690	1.42	.688	.551	.713	.048	.951
CDR-SOB Base Model + Longitudinal DELTA Score Covariates												
Longitudinal DELTA Effects	+m-DELTA + c-DELTA	Fixed	10	11967	5312.6	11987	1.35	.704	.357	.814	.330	.660
	+c-DELTA	Random	11	11805	161.4	11827	1.44	.683	.304	.841	.201	.793
FAQ Model												
FAQ Base Model	Unconditional Means	–	3	35616	–	35622	43.54	–	12.52	–	–	–
	+Time	Fixed + Random	5	33825	1790.4	33835	36.21	.168	4.80	.617	4.89	–
	+Age/Gender/Education	Fixed	8	31355	2470.1	31371	34.99	.196	4.86	.612	5.06	.000
FAQ Base Model + Baseline DELTA Score Covariates												
Baseline DELTA Effects	+BL DELTA Score	Fixed	9	26340	5014.7	26359	18.75	.569	4.94	.605	4.29	.120
	+BL DELTA \times Time	Fixed + Random	11	25719	621.3	25741	19.19	.559	4.71	.624	.42	.914
FAQ Base Model + Longitudinal DELTA Score Covariates												
Longitudinal DELTA Effects	+m-DELTA + c-DELTA	Fixed	10	22052	9303.8	22072	17.84	.590	3.85	.692	1.81	.629
	+c-DELTA	Random	11	21910	142.0	21932	18.84	.567	3.03	.758	1.36	.721

–2LL, –2 Log Likelihood; AIC, Akaike’s Information Criterion; BL, baseline; CDR-SOB, Clinical Dementia Rating Sum of Boxes score; DELTA, Discrepancy-based Evidence for Loss of Thinking Abilities score; c-DELTA, mean-centered DELTA score; FAQ, Functional Activities Questionnaire; m-DELTA, mean DELTA score across time points.

^a All stepwise model additions statistically improved model fit ($p < .0001$).

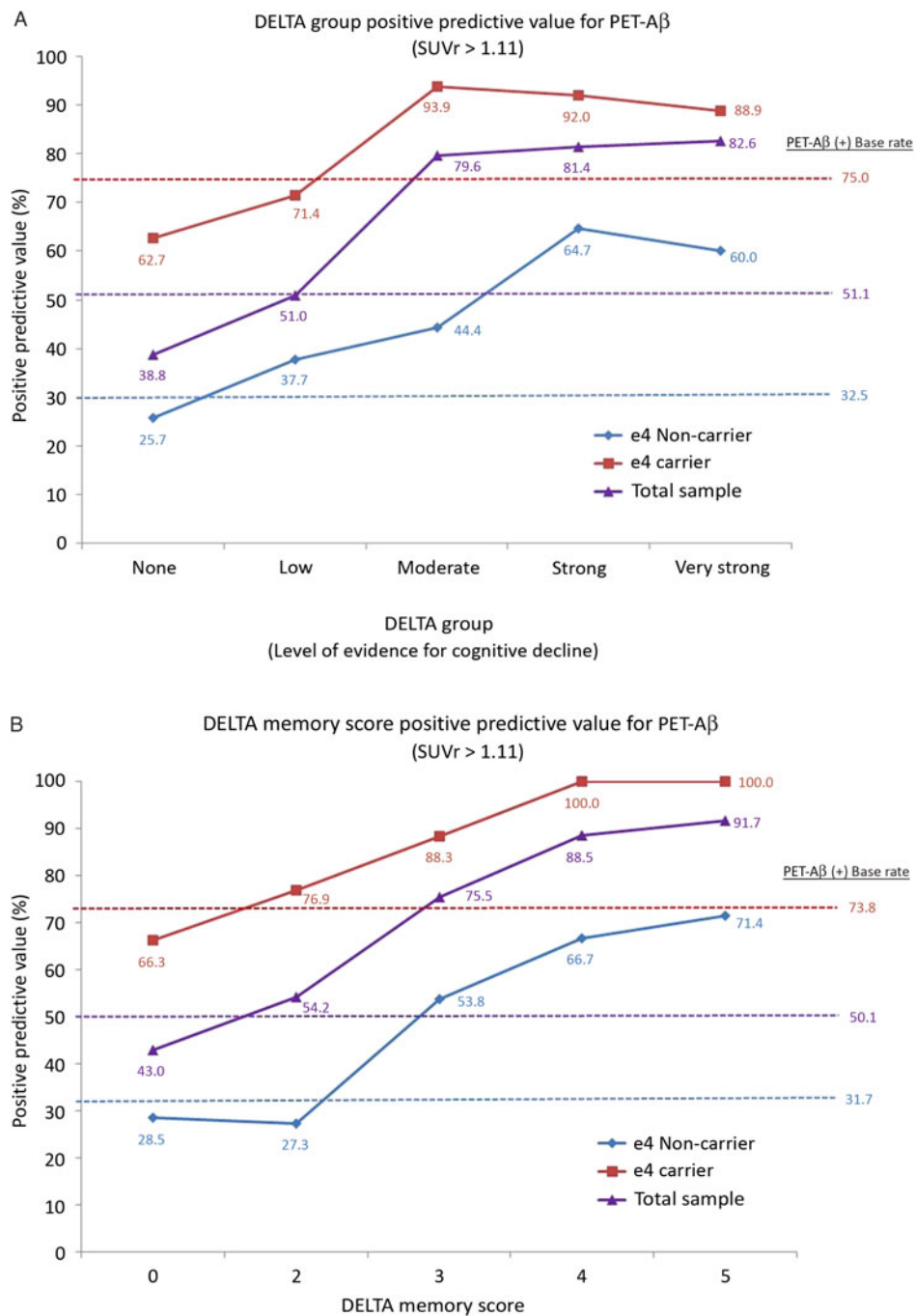


Fig. 2. (A–B): Positive predictive value for PET-A β (SUVr > 1.11) across each DELTA group (A) and stratified by DELTA Memory score (B). Separate lines represent APOE e4 status (carriers vs. noncarriers) and the total sample, with base rates of PET-A β positivity provided for each group (dotted lines). NOTE: No participants obtained a DELTA Memory score of “1” (see Table 2 for criteria).

APOEe4 carriers with CSF-AD biomarkers ($n = 444$), 340 (76.6%) were CSF-AD(+). PPV of the DELTA scores increased from 52.7% in the “No Evidence” group to 95.1% in the “Strong Evidence” group (58/61 participants) and 96.6% in the “Very Strong Evidence” group (28/29 participants). We again observed stronger relationships based on the Memory subscore in the APOEe4 carrier group, such that a Memory scores of 4 (32/32 participants) and 5 (36/36 participants) had 100% PPV.

Automated Scoring Program

An automated scoring program for calculating DELTA scores is provided for free use in Supplemental program.

DISCUSSION

We set out to develop and validate a novel approach for characterizing and quantifying evidence for cognitive decline

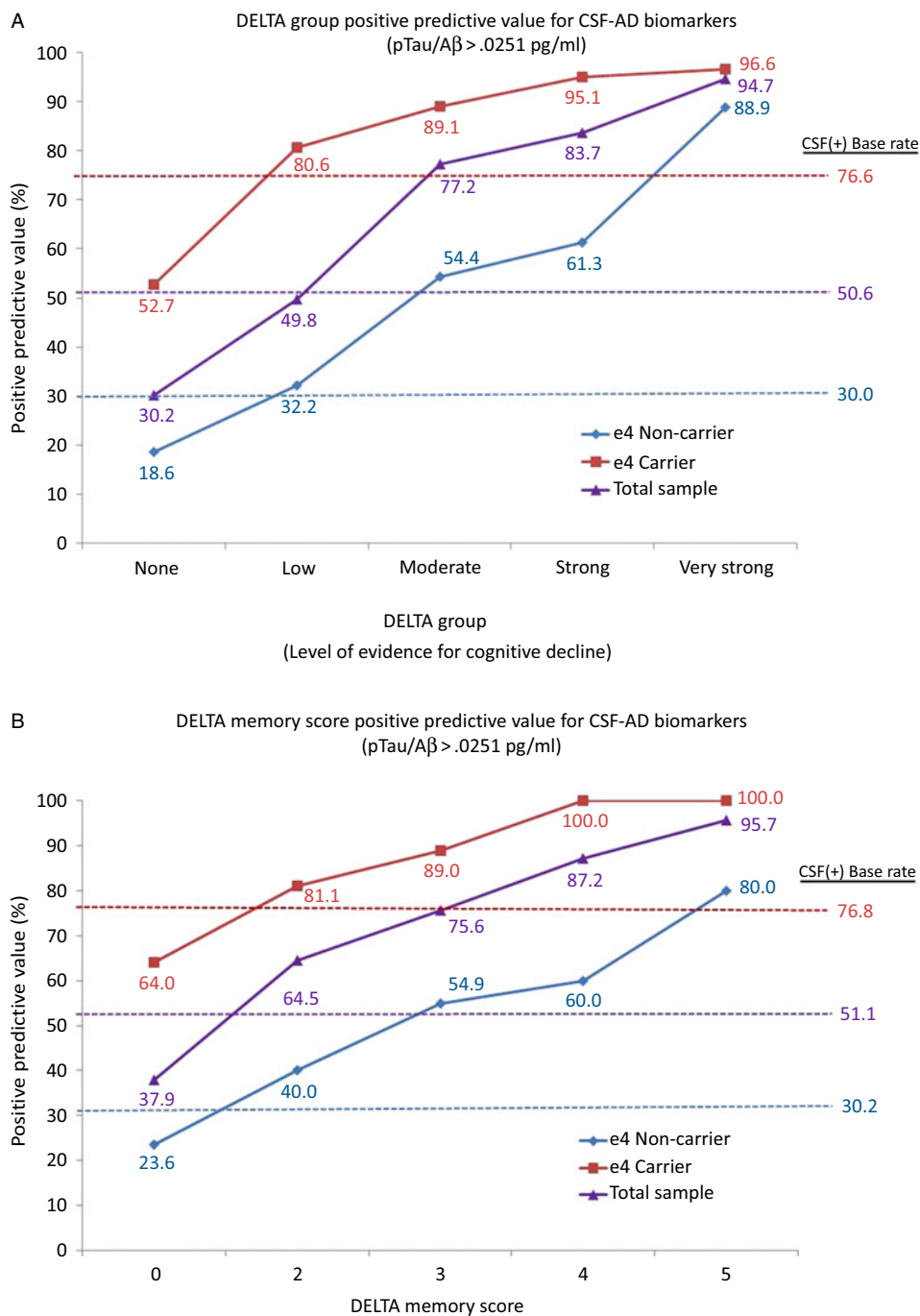


Fig. 3. (A–B): Positive predictive value for CSF-AD biomarkers (pTau/Aβ ratio > .0251 pg/ml) across each DELTA group (A) and stratified by DELTA Memory score (B). Separate lines represent APOE e4 status (carriers vs. noncarriers) and the total sample, with base rates of CSF-AD positivity provided for each group. NOTE: No participants obtained a DELTA Memory score of “1” (see Table 2 for criteria).

based on normative reference methods, which we termed the DELTA score. The DELTA score does not replace existing methods for assessing within-person longitudinal change (e.g., reliable change indices, standardized regression-based change). Novel aspects of the DELTA score and specific considerations for appropriate use are outlined in supplemental material. The approach was rooted in principles of low-score-base-rates aggregated across a relatively comprehensive neuropsychological test battery that evaluated

components of memory, executive function, and language. This is similar conceptually to prior work (Bondi et al., 2014; Jak et al., 2009; Oltra-Cucarella et al., 2018) but differs in that the DELTA Score is continuous compared to typical dichotomization of “impaired” versus “unimpaired” status. We note that while the DELTA scores themselves have somewhat limited variability (0–15 overall, 0–5 per domain), the automated scoring program also produces the continuous standardized discrepancy scores for each test, which can be

flexibly applied in research and clinical settings to fit individual needs.

Prior work using the ADNI cohort separately developed continuous composite scores for memory and executive function, which demonstrated improved prediction of cognitive decline and association with neuroimaging/CSF biomarker outcomes compared to single test scores (Crane et al., 2012; Gibbons et al., 2012). A key finding of these composite approaches was the ability to detect meaningful clinical changes with smaller sample sizes than would be required if using individual tests. This has significant implications when designing clinical trials with cognitive outcomes. Therefore, the DELTA approach may enhance usefulness in this regard given that it is based on multiple test scores and includes multiple cognitive domains, though further research and validation are required.

Jak et al. (2009) showed that “impairment” classification methods with the greatest stability and fewest instances of reversion also likely have the lowest sensitivity to true cognitive impairment (i.e., single-score approaches). This highlights important concepts: (1) progression/reversion/stability rates only matter if the clinician is confident in the diagnostic classification in the first place, (2) diagnostic sensitivity and specificity is a direct function of the strictness of the criteria for determining “impairment” (Iverson & Brooks, 2011). Relatedly, reliance on interpretation of single test scores obtained from a larger test battery may increase risk for both false positives and false negatives. This has been demonstrated regularly in the “multivariate base rate” literature (Binder et al., 2009; Brooks & Iverson, 2010; Houck et al., 2019) and the concept held true in our study as well. For example, if a clinician defined “memory impairment” based on a score <7th percentile of a normative reference group (or $z < -1.5$ in a normal distribution), they are accepting a 7% false positive rate. However, the false positive rate increases exponentially when more test scores are available. Almost 20% of our study’s RCN group would qualify as memory impaired (i.e., one in five scored <7th%ile on either their AVLT or LM delayed recall). In contrast, 97% of our robust normal controls had a DELTA Memory score of 0 thereby illustrating that interpretive approaches like the DELTA score alleviate such problems by taking these concepts into account.

As discussed, there are several strengths of using composite scores derived from comprehensive evaluations for characterizing cognitive status, but there are also practical considerations. Incorporating more tests increases the length of assessments. Other composite scores derived from ADNI data (Crane et al., 2012; Gibbons et al., 2012) included four tests underlying a memory composite [Rey AVLT, WMS-R LM, the Alzheimer’s Disease Assessment Schedule, and MMSE] and five tests for the executive function composite [category fluency (both animals and vegetables), Trail Making Test, Digit Span, WAIS-R Digit-Symbol, Clock Drawing]. The DELTA score in this study comprises a battery of six total tests covering three domains, which we estimate would take 35–40 min. This offers practical advantages

potentially more readily integrated into modern medical settings that emphasize multidisciplinary and time-efficient patient visits. The DELTA score’s high PPV for both PET-A β and CSF-AD biomarker status (+ or –) highlights a potential future application for efficiently identifying (or ruling out) presumably related (or unrelated) disease states for clinical trial enrollment.

Individual neuropsychological tests often have suboptimal test–retest reliability, and therefore scores fluctuate (both higher and lower) due to factors unrelated to the disease process (Brooks et al., 2011). Using normative reference groups demographically and/or intellectually dissimilar to an individual patient also heightens risk for misclassifying cognitive decline (Iverson & Brooks, 2011). Clinicians must be wary of “red herrings” in the form of cognitive test score variability unrelated to disease state. Reducing this phenomenon requires development of more reliable and culturally appropriate measures, and/or classifying cognitive function using multiple test scores in conjunction with low-score-base-rate concepts.

Neuropsychologists uniquely appreciate these concepts and, unsurprisingly, have spearheaded modern approaches for classifying cognitive impairment. However, even the more methodologically rigorous classification criteria often reduce samples to either “impaired” or “unimpaired” status and then characterize by the type of impairment (combinations of single vs. multiple domain and amnesic vs. non-amnesic labels). Dichotomizing cognitive status may contribute to mixed findings regarding clinical progression, reversion, or stability (Pandya, Clem, Silva, & Woon, 2016). Variability in progression, reversion, and stability rates across studies also likely reflects inconsistent definitions for impairment and the number of parameters used for classifying participants (Edmonds et al., 2015; Jak et al., 2009; Thomas et al., 2019).

LIMITATIONS

Multiple limitations of this initial validation coincide with necessary future research outlined below. The current DELTA score was derived from an exclusively white/Caucasian sample that is highly educated. Not every participant contributed data for all follow-up assessment points, likely resulting in survivor bias. Some participants contributed data inconsistently (e.g., BL, Y1, and Y3 but not Y2, Y4, and Y5), which could bias longitudinal frequency rate statistics. Advanced biomarkers were available only on a subset of the total study sample that seemingly was enriched for APOE ϵ 4 carriers (about 40% of those with biomarker data); therefore, PPVs using the total study sample may overestimate general population risk. Age, gender, years of education, and word-reading ability collectively explained only 5–13% of the variance in predicted test scores, suggesting several unmeasured and potentially important factors that could improve the models. Exploration of nonlinear and/or non-mean regression (e.g., quantile regression) when examining the roles of age, education, word-reading, etc. and

accounting for variability in residuals across the spectrum of these variables may further improve premonitory score predictions (Sherwood, Zhou, Weintraub, & Lang, 2016). Sample size was relatively small for certain DELTA groups and associated data should be interpreted cautiously, while further research may refine the cutoff scores associated with a given “level of evidence” for decline group. Lastly, no participants in the study obtained a DELTA Memory score of “1.” Replication in other large samples will help refine scoring criteria, if necessary.

FUTURE DEVELOPMENT AND EXPANSION OF DELTA METHODS

We demonstrated a proof of concept for a novel approach to characterizing evidence for cognitive decline. As with any pilot endeavor, there are many opportunities for expansion and improvement. We propose several ideas that we hope will guide researchers and clinicians in independent replication and validation efforts, and help promote clinical translation.

- Replicate this work in multicultural samples.
- Expand predictors in the regression equations to better explain cognitive test scores. The automated scoring program contains empty fields for “VARIABLE #5” and “VARIABLE #6”, so other researchers can easily adapt the scoring program using new data and novel predictor variables.
- Incorporate tests from additional cognitive domains.
- Validate the DELTA approach using different neuropsychological tests than those used in the present study due to convenience of the ADNI sample.
- We envision opportunities for identifying clinically relevant “profiles” based on patterns of domain-specific DELTA scores. Analogous to “A/T/N” classifications for biomarker evidence of amyloid, tau, and neurodegeneration, we propose something like “M/E/L” for neuropsychological evidence of memory, executive, and language decline using DELTA methodology. We anticipate diverse opinions regarding which cognitive domains to add and which test scores qualify for a given domain.
- Evaluate use of DELTA scores in clinical trials using cognitive outcomes.
- Apply similar methodology for developing a “mood” score and a “behavior” score that could be used in conjunction with the cognitive DELTA score and biomarker panels. This could more precisely characterize clinical syndromes with prominent noncognitive features (e.g., FTD syndromes).

CONCLUSIONS

We present data supporting the initial development and validation of a discrepancy-based test score metric, called the DELTA score, for characterizing the level of evidence for cognitive decline. Higher initial DELTA scores predicted faster rates of functional decline and longitudinal changes in DELTA scores coincided with changes in functional questionnaire scores. Greater evidence for cognitive decline

predicted AD biomarker status, particularly for APOEε4 carriers. Future work should expand the DELTA score to different populations, include additional cognitive domains, and evaluate how domain-specific score patterns align with neurodegenerative disease biomarkers.

ACKNOWLEDGMENTS

This work was supported by an Alzheimer’s Association grant (KRT; AARF-17-528918). The contents of this paper do not represent the views of the Department of Veterans Affairs or the United States Government. 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). 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.

CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <https://doi.org/10.1017/S1355617719001346>

REFERENCES

- Arbuthnott, K. & Frank, J. (2000). Trail making test, part B as a measure of executive control: Validation using a set-switching

- paradigm. *Journal of Clinical and Experimental Neuropsychology*, 22(4), 518–528. doi: [10.1076/1380-3395\(200008\)22:4;1-0;FT518](https://doi.org/10.1076/1380-3395(200008)22:4;1-0;FT518)
- Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J., & Johnson, V.E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. doi: [10.1038/s41562-017-0189-z](https://doi.org/10.1038/s41562-017-0189-z)
- 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.
- Bondi, M.W., Edmonds, E.C., Jak, A.J., Clark, L.R., Delano-Wood, L., McDonald, C.R., Naton, D.A., Libon, D.J., Au, R., Galasko, D., & Salmon, D.P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *Journal of Alzheimer's Disease*, 42(1), 275–289.
- Bondi, M.W. & Smith, G.E. (2014). Mild cognitive impairment: A concept and diagnostic entity in need of input from neuropsychology. *Journal of the International Neuropsychological Society*, 20(2), 129–134.
- Boyle, P.A., Yu, L., Wilson, R.S., Leurgans, S.E., Schneider, J.A., & Bennett, D.A. (2018). Person-specific contribution of neuropathologies to cognitive loss in old age. *Annals of Neurology*, 83(1), 74–83.
- Brooks, B.L. & Iverson, G.L. (2010). Comparing actual to estimated base rates of “abnormal” scores on neuropsychological test batteries: Implications for interpretation. *Archives of Clinical Neuropsychology*, 25(1), 14–21.
- Brooks, B.L., Iverson, G.L., & White, T. (2009). Advanced interpretation of the Neuropsychological Assessment Battery with older adults: Base rate analyses, discrepancy scores, and interpreting change. *Archives of Clinical Neuropsychology*, 24(7), 647–657.
- Brooks, B.L., Sherman, E.M., Iverson, G.L., Slick, D.J., & Strauss, E. (2011). Psychometric foundations for the interpretation of neuropsychological test results. In *The Little Black Book of Neuropsychology*, (pp. 893–922). Boston, MA: Springer.
- Casaletto, K.B., Marx, G., Dutt, S., Neuhaus, J., Saloner, R., Kritikos, L., Miller, B., & Kramer, J.H. (2017). Is “Learning” episodic memory? Distinct cognitive and neuroanatomic correlates of immediate recall during learning trials in neurologically normal aging and neurodegenerative cohorts. *Neuropsychologia*, 102, 19–28.
- Crane, P.K., Carle, A., Gibbons, L.E., Insel, P., Mackin, R.S., Gross, A., Jones, R.N., Mukherjee, S., Curtis, S. M., Harvey, D., & Weiner, M. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, 6(4), 502–516.
- Crawford, J.R., Moore, J.W., & Cameron, I.M. (1992). Verbal fluency: A NART-based equation for the estimation of premorbid performance. *British Journal of Clinical Psychology*, 31(3), 327–329.
- De Meyer, G., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., De Deyn, P.P., Coart, E., Hansson, O., Minthon, L., Zetterberg, H., & Blennow, K. (2010). Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Archives of Neurology*, 67(8), 949–956. doi: [10.1001/archneurol.2010.179](https://doi.org/10.1001/archneurol.2010.179)
- Dotson, V.M., Kitner-Triolo, M.H., Evans, M.K., & Zonderman, A.B. (2009). Effects of race and socioeconomic status on the relative influence of education and literacy on cognitive functioning. *Journal of the International Neuropsychological Society*, 15(4), 580–589.
- Duff, K., Chelune, G.J., & Dennett, K. (2011). Predicting estimates of premorbid memory functioning: Validation in a dementia sample. *Archives of Clinical Neuropsychology*, 26(8), 701–705.
- Duff, K., Dalley, B., Suhrie, K.R., & Hammers, D.B. (2018). Predicting premorbid scores on the repeatable battery for the assessment of neuropsychological status and their validation in an elderly sample. *Archives of Clinical Neuropsychology*, 34(3), 395–402.
- Edmonds, E.C., Delano-Wood, L., Clark, L.R., Jak, A.J., Naton, D.A., McDonald, C.R., Libon, D.J., Au, R., Galasko, D., Salmon, D.P., & Bondi, M.W. (2015). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's & Dementia*, 11(4), 415–424.
- Eppig, J.S., Edmonds, E.C., Campbell, L., Sanderson-Cimino, M., Delano-Wood, L., Bondi, M.W., & Alzheimer's Disease Neuroimaging, I. (2017). Statistically derived subtypes and associations with cerebrospinal fluid and genetic biomarkers in mild cognitive impairment: A latent profile analysis. *Journal of the International Neuropsychological Society*, 23(7), 564–576. doi: [10.1017/S135561771700039X](https://doi.org/10.1017/S135561771700039X)
- Gibbons, L.E., Carle, A.C., Mackin, R.S., Harvey, D., Mukherjee, S., Insel, P., Curtis, S.M., Mungas, D., Crane, P.K., & Alzheimer's Disease Neuroimaging Initiative. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior*, 6(4), 517–527.
- Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J.Q., Bittner, T., Lifke, V., Corradini, V., Eichenlaub, U., Batrla, R., & Buck, K. (2018). CSF biomarkers of Alzheimer's disease concord with amyloid- β PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts. *Alzheimer's & Dementia*, 14(11), 1470–1481.
- Houck, Z.M., Asken, B.M., Bauer, R.M., Kontos, A.P., McCrea, M.A., McAllister, T.W., Broglio, S.P., Clugston, J.R., & Care Consortium Investigators. (2019). Multivariate base rates of low scores and reliable decline on ImPACT in healthy collegiate athletes using Care Consortium norms. *Journal of the International Neuropsychological Society*, 25(9), 961–971.
- Ioannidis, J.P.A. (2018). The proposal to lower p value thresholds to .005. *JAMA*, 319(14), 1429–1430. doi: [10.1001/jama.2018.1536](https://doi.org/10.1001/jama.2018.1536)
- Iverson, G.L. & Brooks, B.L. (2011). Improving accuracy for identifying cognitive impairment. In M.R. Schoenberg & J.G. Scott (Eds.), *The Little Black Book of Neuropsychology*, (pp. 923–950). Boston, MA: Springer.

- Jack, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeblerlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., & Liu, E. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562.
- 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. *The American Journal of Geriatric Psychiatry*, *17*(5), 368–375.
- Jak, A.J., Preis, S.R., Beiser, A.S., Seshadri, S., Wolf, P.A., Bondi, M.W., & Au, R. (2016). Neuropsychological criteria for mild cognitive impairment and dementia risk in the Framingham Heart Study. *Journal of the International Neuropsychological Society*, *22*(9), 937–943.
- James, B.D., Wilson, R.S., Boyle, P.A., Trojanowski, J.Q., Bennett, D.A., & Schneider, J.A. (2016). TDP-43 stage, mixed pathologies, and clinical Alzheimer's-type dementia. *Brain*, *139*(11), 2983–2993.
- Landau, S., Thomas, B., Thurfjell, L., Schmidt, M., Margolin, R., Mintun, M., Pontecorvo, M., Baker, S.L., Jagust, W.J., & Alzheimer's Disease Neuroimaging Initiative. (2014). Amyloid PET imaging in Alzheimer's disease: A comparison of three radiotracers. *European Journal of Nuclear Medicine and Molecular Imaging*, *41*(7), 1398–1407.
- Manly, J.J. & Echemendia, R.J. (2007). Race-specific norms: Using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology*, *22*(3), 319–325. doi: [10.1016/j.acn.2007.01.006](https://doi.org/10.1016/j.acn.2007.01.006)
- McGurn, B., Starr, J., Topfer, J., Pattie, A., Whiteman, M., Lemmon, H.A., Whalley, L.J., & Deary, I. (2004). Pronunciation of irregular words is preserved in dementia, validating premorbid IQ estimation. *Neurology*, *62*(7), 1184–1186.
- Morris, J.C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, *43*(11), 2412–2414.
- Mortamais, M., Ash, J.A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., Pose, C., Albalá, B., Ropacki, M., Ritchie, C.W., & Ritchie, K. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's & Dementia*, *13*(4), 468–492. doi: [10.1016/j.jalz.2016.06.2365](https://doi.org/10.1016/j.jalz.2016.06.2365)
- Nelson, P.T., Dickson, D.W., Trojanowski, J.Q., Jack, C.R., Boyle, P.A., Arfanakis, K., Rademakers, R., Alafuzoff, I., Attems, J., Brayne, C., Coyle-Gilchrist, I.T., & Schneider, J.A. (2019). Limbic-predominant age-related TDP-43 encephalopathy (LATE): Consensus working group report. *Brain*, *142*(6), 1503–1527. doi: [10.1093/brain/awz099](https://doi.org/10.1093/brain/awz099)
- Oltra-Cucarella, J., Sánchez-SanSegundo, M., Lipnicki, D.M., Sachdev, P.S., Crawford, J.D., Pérez-Vicente, J. A., Cabello-Rodríguez, L., Ferrer-Cascales, R., & Alzheimer's Disease Neuroimaging Initiative. (2018). Using base rate of low scores to identify progression from amnesic mild cognitive impairment to Alzheimer's disease. *Journal of the American Geriatrics Society*, *66*(7), 1360–1366.
- Ossenkoppelle, R., Pijnenburg, Y.A., Perry, D.C., Cohn-Sheehy, B.I., Scheltens, N.M., Vogel, J.W., Kramer, J.H., van der Vlies, A.E., Joie, R.L., Rosen, H.J., & van der Flier, W.M. (2015). The behavioural/dysexecutive variant of Alzheimer's disease: Clinical, neuroimaging and pathological features. *Brain*, *138*(Pt 9), 2732–2749. doi: [10.1093/brain/awv191](https://doi.org/10.1093/brain/awv191)
- Pandya, S.Y., Clem, M.A., Silva, L.M., & Woon, F.L. (2016). Does mild cognitive impairment always lead to dementia? A review. *Journal of the Neurological Sciences*, *369*, 57–62.
- Perry, D.C., Brown, J.A., Possin, K.L., Datta, S., Trujillo, A., Radke, A., Karydas, A., Kornak, J., Sias, A.C., Rabinovici, G.D., & Gorno-Tempini, M.L. (2017). Clinicopathological correlations in behavioural variant frontotemporal dementia. *Brain*, *140*(12), 3329–3345.
- Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, *56*(3), 303–308.
- Pfeffer, R.I., Kurosaki, T.T., Harrah Jr, C.H., Chance, J.M., & Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of Gerontology*, *37*(3), 323–329.
- Phillips, J.S., Da Re, F., Dratch, L., Xie, S.X., Irwin, D.J., McMillan, C.T., Vaishnavi, S.N., Ferrarese, C., Lee, E.B., Shaw, L.M., & Trojanowski, J.Q. (2018). Neocortical origin and progression of gray matter atrophy in nonamnesic Alzheimer's disease. *Neurobiology of Aging*, *63*, 75–87. doi: [10.1016/j.neurobiolaging.2017.11.008](https://doi.org/10.1016/j.neurobiolaging.2017.11.008)
- Rivera Mindt, M., Byrd, D., Saez, P., & Manly, J. (2010). Increasing culturally competent neuropsychological services for ethnic minority populations: A call to action. *The Clinical Neuropsychologist*, *24*(3), 429–453. doi: [10.1080/13854040903058960](https://doi.org/10.1080/13854040903058960)
- Sherwood, B., Zhou, A.X.H., Weintraub, S., & Wang, L. (2016). Using quantile regression to create baseline norms for neuropsychological tests. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *2*, 12–18.
- Thomas, K.R., Eppig, J.S., Weigand, A.J., Edmonds, E.C., Wong, C.G., Jak, A.J., Delano-Wood, L., Galasko, D.R., Salmon, D.P., Edland, S.D., & Bondi, M.W. (2019). Artificially low mild cognitive impairment to normal reversion rate in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *15*(4), 561–569.
- Wennberg, A.M., Whitwell, J.L., Tosakulwong, N., Weigand, S.D., Murray, M.E., Machulda, M.M., Petrucelli, L., Mielke, M.M., Jack Jr, C.R., Knopman, D.S., & Parisi, J.E. (2019). The influence of tau, amyloid, alpha-synuclein, TDP-43, and vascular pathology in clinically normal elderly individuals. *Neurobiology of Aging*, *77*, 26–36.