Alzheimer's & Dementia 5 (2009) 295–306

Development and validation of the Memory Performance Index: Reducing measurement error in recall tests

William R. Shankle*, Tushar Mangrola, Timothy Chan, Junko Hara

Medical Care Corporation, Irvine, CA, USA

Abstract

Background: The Memory Performance Index (MPI) quantifies the pattern of recalled and nonrecalled words of the Consortium to Establish a Registry for Alzheimer's Disease Wordlist (CWL) onto a 0 to 100 scale and distinguishes normal from mild cognitive impairment with 96% to 97% accuracy.

Methods: In group A, 121,481 independently living individuals, 18 to 106 years old, were assessed with the CWL and classified as cognitively impaired (N = 5,971) or normal (N = 115,510). The MPI and CWL immediate free recall (IFR), delayed free recall (DFR), and total free recall (TFR) scores (the outcome measures) were each regressed against predictors of age, gender, race, education, test administration method (in-person or telephone), and wordlist used. Predictor effect sizes (Cohen's f^2) were computed for each outcome. In addition, CWL plus Functional Assessment Staging Tests (FAST) were administered to 441 normal to moderately severely demented (FAST stages 1 to 6) patients (group B). Median MPI scores were tested for significant differences across FAST stage.

Results: For group A, the variance explained by all predictors combined was MPI = 55.0%, IFR = 24.9%, DFR = 23.4%, and TFR = 26.9%. The age effect size on MPI score was large, but it was small on IFR, DFR, and TFR. The other predictors all had negligible (<0.02) or small effect sizes (0.02 to 0.15). For group B, median MPI scores progressively declined across all FAST stages ($P < .0002$).

Conclusions: MPI score progressively declines with increasing dementia severity. Also, MPI score explains 2 to 3 times more variance than total scores, which improves ability to detect treatment effects.

Keywords: Alzheimer's disease; Mild cognitive impairment; Dementia; Accuracy; Normal aging; Episodic memory; Declarative memory; Short-term memory; Hippocampus; Working memory; Correspondence analysis; Receiver operating characteristic; Logistic regression

1. Introduction

1.1. Consortium to Establish a Registry for Alzheimer's Disease Wordlist Recall Test

The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Wordlist Recall Test (CWL) is a standardized, well-validated assessment of immediate free recall (IFR) and delayed free recall (DFR) developed in the 1980s by the National Institute of Aging Alzheimer's Disease Centers [1,2]. The traditional scoring uses a cutoff score based on the number of words recalled during the three learning trials or during the DFR trial, for which adjustment might or might not be made to account for the age of the subject. A study of community-based versus university subject samples showed that normal aging subjects recalled a mean of 19.5 ± 5.0 to 21.4 ± 4.4 words during the three CWL learning trials and recalled a mean of 6.0 ± 2.8 to 7.5 ± 2.0 words on DFR [3]. One of the few studies that included CWL cutoff scores for patients with mild cognitive impairment (MCI) was one involving the Finnish version of the CERAD neuropsychological test battery [4]. In that study, the CERAD battery was administered to patients with normal aging, amnestic MCI (aMCI), and mild dementia caused by Alzheimer's disease (AD). The authors found that the CWL test gave the best discrimination for these patient groups and reported optimal cutoff scores for the sum of the three learning trials (16/30) and for DFR trial (6/10), which gave respective sensitivities of 0.33 and 0.33 for aMCI and of 0.6 and 0.86 for mild AD

*Corresponding author. Tel.: 949-838-0154; Fax: 949-838-0153.
E-mail address: rshankle@mccare.com

1552-5260/09/$ – see front matter © 2009 The Alzheimer's Association. All rights reserved.
doi:10.1016/j.jalz.2008.11.001

dementia, with specificities of 0.93 and 1.0 for normal aging. The authors suggested increasing the number of list words or the time between learning and DFR to further improve detection of aMCI cases.

1.2. Improving the CWL scoring with recall pattern analyses

An alternative way to improve detection of MCI is to use more of the available information in the CWL test. When one considers that there are at least 1 trillion possible patterns for recalling 10 words across four trials, the reliance on total scores for learning and DFR ignores almost all of the available information.

We have previously introduced a mathematical algorithm that measures the pattern of both recalled and nonrecalled words across the four CWL trials and classifies the pattern as cognitively normal or impaired on the basis of a cut point that characterizes sensitivity and specificity levels appropriate to the requirements of the clinical setting [5]. On the basis of nonparametric receiver operating characteristic curve determination of overall accuracy, this algorithm discriminates normal aging from MCI by 96% to 97% and discriminates normal aging from mild dementia with 99% accuracy [5–7]. The inter-rater reliability was 0.83, and the diagnostic validity for patients in Functional Assessment Staging Tests (FAST) stages 1 to 4 has a kappa value of 0.92 ± 0.09 [6].

The algorithm's parameters were originally derived from an analysis of a sample of 471 well-characterized subjects who had no cognitive or functional impairment, had MCI, or had mild dementia [5]. For the majority of the impaired subjects, underlying causes included AD, Lewy body disease, frontotemporal lobe disease, and cerebrovascular disease. A small number of subjects had other dementing disorders.

The score produced by this algorithm measures characteristics that are not captured by total scores of the numbers of correctly recalled words [5]. These characteristics include differential effects on recall difficulty as a function of (1) a word's position in the learning list, (2) the number of times a subject has been exposed to the word, (3) the delay between learning the list and recalling its words, and (4) the patterns of recall across the learning and testing trials that are unique to persons with no cognitive impairment, MCI, or mild dementia. In addition, the algorithm's score also measures the effect of not recalling a word in a given trial. This effect is also influenced by word position, frequency of exposure, and delay between exposure and recall.

For example, Fig. 1 in our previous publication [5] shows that the first (w1) and seventh to tenth words (w7 to w10) in the learning list are easier to recall than words second to sixth (w2 to w6), which is consistent with well-known effects of primacy and recency. Also, for a given list word such as *actor*, immediate recall is hard on trial 1, easier on trial 2, easiest on trial 3, and then is hardest after a several minute delay. Consider a real example of similarly aged subjects who recalled 6 of 10 on IFR trial 3 and 4 of 10 on DFR trial 4. The aforementioned algorithm classified one subject as impaired and the other as normal. The orders in which they recalled list words w1 to w10 on trial 4 were the following:

w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	Classification
0	0	0	0	0	0	2	4	1	3	Impaired
1	0	4	3	0	2	0	0	0	0	Normal

In spite of identical recall totals on trials 3 and 4, the patterns of recall for the normal and impaired subjects differed, with the impaired subject recalling the most recently exposed words from the end of the list and the normal subject recalling words more closely to the order in which they were presented. Howard and Kahana [8] have shown that recall order approximates the order of presented stimuli when the stimuli cannot be easily associated. This low associability among words is the case for the wordlists used in the CWL and in the MCI Screen (MCIS), a web-based implementation of the CWL that uses the aforementioned scoring algorithm.

1.3. Quantifying wordlist recall pattern

Although a classification of the pattern of recalled and not recalled words is useful to discriminate between healthy subjects and those with some underlying cause of progressive memory loss, a quantification of such a pattern offers a more useful and intuitive understanding of overall memory function and might provide more precise measurement of longitudinal change.

The present article presents the Memory Performance Index (MPI), a scale from 0 to 100 with cut point centered at 50, which quantifies a subject's pattern of recalled and not recalled words. Through a monotonic transformation, the MPI scale provides a more useful interpretation of the results produced by the CWL scoring algorithm. The present article also analyzed the amount of variance of the MPI score and of the total numbers of words correctly recalled immediately or after a delay that can be accounted for by typical sources of variability such as age, gender, race, education, test administration method, and test stimuli. For this purpose, a sample of 121,481 long-term care (LTC) insurance applicants aged 18 to 106 years old were analyzed. Finally, the relation between MPI score and dementia severity was analyzed by using a well-characterized clinical sample of 441 cognitively normal to moderately severely demented patients.

2. Methods

2.1. Development of the MPI

The MCIS is a web-based assessment tool that guides examiners to reliably administer the CWL test plus additional measures of executive function [6]. The CWL scoring algorithm has been adopted as a cognitive measure in academic, clinical, disease management, and insurance settings.

The CWL scoring algorithm [5] uses correspondence analysis [9], a technique that creates weighted scores from the subject's full CWL performance profile, which consists of the pattern of recalled and not recalled words across four trials. The methods by which the CWL scoring algorithm was derived are fully described elsewhere [5]. Correspondence analysis produces an optimally weighted combination of values, which are then used in a logistic regression to predict each subject's probability of cognitive impairment.

The value derived from the logistic regression (the logistic regression score) represents a subject's full CWL performance profile as a single value that has an unwieldy range from a negative number to a positive number that is hard to work with. Therefore, the logistic regression score was translated onto a scale, the MPI, of 0 to 100, with 50 making the cut point between impaired (<50) and normal (>50). The MPI scale is constructed as follows:

1. Define the MPI scale to range from 0 to 100, with cut point separating normal from cognitively impaired centered at 50.
2. Transform the 99% confidence interval (CI) of the logistic regression score cut point onto the MPI scale.
3. Transform the logistic regression scores that fall into the cognitively impaired range to have MPI scores between 0 and the lower limit of the 99% CI of the MPI scale cut point.
4. Transform the logistic regression scores that fall into the normal range to have MPI scores between the upper limit of the 99% CI of the MPI scale cut point and 100.
5. Transform the logistic regression scores that fall within the 99% CI of the cut point to fall within the 99% CI of the MPI cut point and classify these cases as borderline.

Figure 2 summarizes the formulae for the 99% CI of the MPI cut point and for computing a given subject's MPI score from their logistic regression score.

2.2. Measurement characteristics of the MPI

To characterize memory performance variability in normal aging, cognitive impairment, and differing levels of dementia severity, covariates (predictor variables) were identified and tested for their influence on the outcome variables of MPI score and of total scores of the numbers of correctly recalled words during IFR and DFR trials. Two different data samples, groups A and B, were used to characterize memory performance variability.

2.2.1. Outcome and predictor variables

The outcome variables were the MPI score and the numbers of correctly recalled words during three IFR trials, the one DFR trial, and all four trials combined (TFR).

Potential predictors accounting for variability in these outcome measures were age, gender, race, education, test administration method (in-person versus telephone), and wordlist used to test a subject. Cubic and linear spline analyses (Stata 10.0, mkspline; Stata Corporation, College Station, TX) were used to create ordinal variables for age and education without losing explanatory power on the outcome variables. The optimal age groups identified were 18 to 51, >51 to 65, >65 to 77, and >77 years. The optimal education levels identified were 0 to 12, >12 to 16, and >16 years of education.

2.3. Data sample: Group A

The first sample (group A) consisted of an initial cohort of 125,238 LTC insurance applicants aged 18 to 106 years who were all living independently in the community. The initial assessment consisted of applicant interview to identify risk factors for AD and related disorders (ADRD), impaired instrumental or basic activities of daily living, behavioral problems, use of a cholinesterase inhibitor, memantine, antipsychotic, or antidepressant medication, prior diagnosis of dementia, cognitive impairment, ADRD, cerebrovascular disease, cardiovascular disease, cancer, depression, bipolar disorder, psychosis, schizophrenia, and other psychiatric disorders. Because this assessment is required before issuing an LTC policy, no applicants refused. In cases in which there was either a high risk or a suspicion of a high risk for ADRD, medical records were also reviewed. Three thousand seven hundred fifty-seven (3%) applicants were deemed to have high risk for cognitive impairment, functional impairment, or dementia and were excluded from further assessment. The criteria for exclusion depended on risk factors noted above. The remaining 121,481 applicants were tested with the MCIS. When an applicant took the MCIS more than once, we only used the initial MCIS result to avoid test-retest bias. Figure 1 shows the sample selection process for group A.

2.4. Classification of normal versus cognitive impairment

The CWL scoring algorithm has an overall classification accuracy (area under the receiver operating characteristic [ROC] curve) of 96% to 97% for discriminating between cognitively normal aging and progressively declining MCI in three independent studies of a total of 492 subjects evaluated in primary care and specialty settings in the United States and Japan [5–7]. We used a cut point on the ROC curve corresponding to a sensitivity of 96% for impaired and a specificity of 88% for normal to select cognitively normal cases for additional analysis to minimize biasing normally classified subjects with those classified as impaired. Of the 121,481 applicants assessed, 115,510 were classified as normal with the specified cut point criteria.

2.5. Group A sample characteristics

Table 1 shows the age group distributions of cases for each predictor variable (gender, race, education level, test method, and wordlist). We computed the Cramer's phi (φ_c) statistic

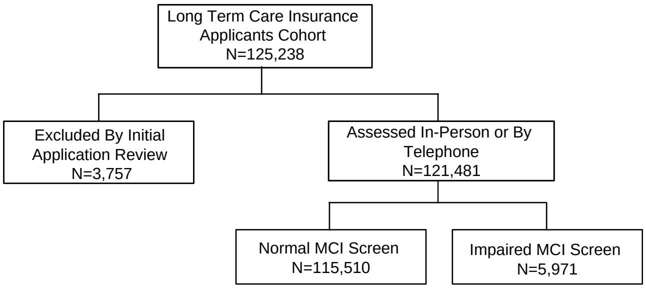
Long Term Care Insurance Applicants Cohort N=125,238
Excluded By Initial Application Review N=3,757
Assessed In-Person or By Telephone N=121,481
Normal MCI Screen N=115,510
Impaired MCI Screen N=5,971

Fig. 1. Sample selection diagram for the LTC insurance applicant sample (group A).

for each category of a predictor variable within each of the four age groups as well as for all ages combined. Cramer's φ_c measures the effect size of each cell in terms of its χ^2 value normalized by the total sample size (N) of each predictor variable and by its number of categories (k) [10]. Cells with effect sizes that are medium (0.15 to 0.35) or large (>0.35) have superscripts of m and l, respectively. The formula for Cramer's phi (φ_c) is the following:

$$\phi_c \equiv \sqrt{\frac{\chi^2}{N(k-1)}}$$

2.6. Data sample: Group B

The other sample (group B) consisted of 441 patients evaluated for ADRD at a primary care setting with special interest in ADRD (N = 138) or a dementia specialty clinic (N = 303). The primary care data were originally collected as part of the Hancock County Aging Project in Ellsworth, Maine. Further details of the sample description are discussed elsewhere [6]. The dementia data samples were collected at the Shankle Clinic in Irvine, California, which specializes in ADRD. All patients had a standardized diagnostic assessment. Laboratory studies included those relevant to identifying dementing disorders (B_{12}, folate, homocysteine, thyroid-stimulating hormone, free thyroxine, erythrocyte sedimentation rate, antinuclear antibody titer, complete blood count with differential, chemistry panel, and apolipoprotein E genotype). Brain imaging consisted of magnetic resonance imaging and, when indicated, a fluorodeoxyglucose positron emission tomography scan. Medical, family, and lifestyle history relevant to ADRD differential diagnosis was carefully reviewed. These data were then used to diagnose AD versus non-AD etiologies in conformance with National Institute of Neurological and Communicative Diseases and Stroke/Alzheimer's Disease and Related Disorders Association diagnostic criteria [11–13]. The data for the 138 primary care clinic cases were reviewed, and the ADRD diagnoses were confirmed by the Shankle Clinic neurologist to improve ADRD diagnostic accuracy and reliability.

Each patient was reassessed every 3 to 6 months. The following measures were collected at each reassessment: (1) cognition by using the MCIS plus several measures of executive function (subtests of the Delis Kaplan battery, CERAD Trails tests, and CERAD letter fluency); (2) dementia severity and functional capacity by using the FAST [14,15]; (3) review and adjustment of medications; (4) review of medical problems; and (5) review of ADRD-related comorbidities. There was a total of 2,416 reassessments for the 441 patients, who were followed for up to 5 years. These patients ranged from cognitively normal to moderately severely demented (Table 2).

2.7. Sample characteristics

Table 2 characterizes the group B sample by FAST stage. Effect sizes for MPI score, age, and years of education were computed by using Hedges' g statistic [16], and those for female gender were computed as odds ratios. For all effect size measures, FAST stages 2 to 6 were compared with FAST stage 1 (a normal aging sample). Hedges' g statistic is the following:

$$\hat{g} \equiv \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{N_{total}-2}}} \times \left(1-\left(\frac{3}{4(n_1+n_2)-9}\right)\right)$$

2.8. Measuring the magnitude of effect of each predictor on the outcome measures

The entire group A sample was examined to measure the variability in the scores of the outcome measures that could be accounted for by the predictor variables, which are defined below.

2.8.1. Outcome and predictor variables

The outcome measures were the MPI score, IFR, DFR, and TFR. The predictor variables examined were age group, race (white, Asian, Hispanic, black, and Native [Pacific Islander/American Indian/other native group]), gender, educational level, test method (in-person, telephone), and wordlist used.

2.9. Proportion of variance explained and predictor effect size

To examine the proportion of each outcome measure's variance explained by the full set of predictor variables, we analyzed the entire group A sample of cognitively normal and impaired subjects (N = 121,481), as well as the sample of those classified as cognitively normal (N = 115,510). The predictor, race, was not a required datum and was often omitted and had a smaller sample size of 8,033 subjects. Stepwise regression (Stata 10.0, sw regress; Stata Corporation) of the full set of predictors against each outcome measure removed no predictors. For each outcome measure, multiple linear regression (Stata 10.0 regress; Stata Corporation) was used to determine the proportion of variance explained by the full set of predictor variables.

To graphically display differences in the proportion of variance explained for the four outcome measures, we performed a piecewise linear regression of the four age groups listed in Table 1 on each outcome (N = 121,481). The 95% confidence

To illustrate the computation of the MPI score and scale, assume that optimal CWL scores for *impaired* individuals are less than those for *normal* individuals. Let:

X_{min} = the theoretical minimum possible optimal CWL score

X_{max} = the theoretical maximum possible optimal CWL score

X_c = the optimal CWL score cut-point separating *normal* from *cognitive impairment*

X_{ll99} = the lower limit of the 99% confidence interval (CI) of the optimal CWL score cut-point separating normal from impaired individuals.

X_{ul99} = the upper limit of the 99% CI of the optimal CWL score cut-point separating normal from impaired individuals.

Y_{min} = 0, the theoretical minimum possible MPI score

Y_{max} = 100, the theoretical maximum possible MPI score

Y_c = 50, the pre-defined MPI score cut-point separating *normal* from *cognitive impairment*

Y_{ll99} = the lower limit of the 99% CI of the MPI score cut-point

Y_{ul99} = the upper limit of the 99% CI of the MPI score cut-point

The lower and upper limits of the 99% CI of the MPI cut-point are therefore:

$$Y_{ll99} \equiv 50 - \left[100 \times \frac{(X_c - X_{ll99})}{X_{max} - X_{min}}\right]$$

$$Y_{ul99} \equiv 50 + \left[100 \times \frac{(X_{ul99} - X_c)}{X_{max} - X_{min}}\right]$$

The following formula transforms an individual's optimal CWL score, X_i, to their MPI score, Y_i:

$$If\ X_i\ is\ normal: Y_i \equiv Y_{ul99} + (Y_{max} - Y_{ul99}) \times \frac{(X_i - X_{ul99})}{(X_{max} - X_{ul99})}$$

$$If\ X_i\ is\ impaired: Y_i \equiv Y_{ll99} \times \frac{(X_i - X_{min})}{(X_{ll99} - X_{min})}$$

$$If\ X_i\ is\ borderline: Y_i \equiv Y_c + \frac{100 \times (X_i - X_c)}{(X_{max} - X_{min})}$$

Fig. 2. Method for transforming the logistic regression scale and score into the MPI scale and score. The first two equations define the 99% CI for the MPI cut point centered at 50. The remaining equations define the logistic regression score transformation to the MPI score for a given subject, depending on their cognitive classification.

bands were then computed and plotted with four piecewise linear regressions and the age distribution of the data for each outcome measure (Stata10.0 scatter lfit; Stata Corporation). The resulting graphs are shown in Fig. 3.

To measure the magnitude of a given predictor's effect, we first regressed each outcome measure against all predictors and then repeated the regression with the target predictor removed. The predictor, test method, was systematically biased in favor of telephone testing for younger subjects and in-person testing for older subjects (Table 1). We controlled for this bias by randomly sampling 330 subjects within each combination of age group and test method (330 was approximately one third of the smallest number of subjects tested by either method in any age group). We used bootstrap random sampling with replacement, repeated 100 times, to obtain estimates of the effect size for the predictor, test method (Stata 10.0, bootstrap).

Cohen's f^2 was used to measure the predictor's effect size in explaining the variance of each outcome measure. Cohen's f^2 expresses the variance explained by the predictor in terms of the total variance not explained by the full set of predictors [10]. The formula for Cohen's f^2 is the following:

$$f^2 \equiv \frac{\left(R^2_{AB} - R^2_A\right)}{1 - R^2_{AB}},$$

where $R^2{}_{\mathrm{AB}}$ is the proportion of variance explained by the full set of predictors, and $R^2{}_{\mathrm{A}}$ is the proportion explained after removing the predictor whose effect size one wishes to measure. f^2 effect sizes of 0.02, 0.15, and 0.35 are considered small, medium, and large, respectively. Table 3 reports the effect size that each predictor has on each outcome (MPI, IFR, DFR, and TFR). These effect sizes are reported for the entire group A sample as well as for the subset of cases classified as normal. The last row of Table 3 also reports the proportion of variance explained by the full set of predictors for each outcome measure.

2.10. MPI score in relation to dementia severity

Dementia severity was assessed for all 441 cases from group B by using the FAST. The FAST staging is a well-standardized measure of the course of cognitively related

Table 1
Age group distributions of predictor variables for the group A sample [numbers of cases (% in age group)]

Group A sample	Age group (y)				
	18–50	51–64	65–76	77–106	Total
Years of education					
0–12	1,095 (9.2)	5,201 (10.7)	6,501 (13.5)	2,489 (19.5)	15,286 (12.6)
13–16	5,705 (47.8)	22,372 (46.2)	24,159 (50)	6,355 (49.7)	58,591 (48.2)
17–30	5,146 (43.1)	20,888 (43.1)	17,615 (36.5)	3,955 (30.9)	47,604 (39.2)
Total	11,946 (100)	48,461 (100)	48,275 (100)	12,799 (100)	121,481 (100)
Gender					
Female	6,743 (56.5)	28,060 (57.9)	25,703 (53.2)	7,275 (56.8)	67,781 (55.8)
Male	5,203 (43.6)	20,401 (42.1)	22,572 (46.8)	5,524 (43.2)	53,700 (44.2)
Total	11,946 (100)	48,461 (100)	48,275 (100)	12,799 (100)	121,481 (100)
Race					
Asian	44 (0.4)	153 (0.3)	72 (0.2)	40 (0.3)	309 (0.3)
Black	48 (0.4)	78 (0.2)	85 (0.2)	57 (0.5)	268 (0.2)
Hispanic	61 (0.5)	75 (0.2)	116 (0.2)	57 (0.5)	309 (0.3)
Native	13 (0.1)	15 (0)	11 (0)	14 (0.1)	53 (0)
White	1,157 (9.7)	2,367 (4.9)	2,016 (4.2)	1,551 (12.1)^{m}	7,091 (5.8)
NOS	10,623 (88.9)	45,773 (94.5)	45,975 (95.2)	11,080 (86.6)	113,451 (93.4)
Total	11,946 (100)	48,461 (100)	48,275 (100)	12,799 (100)	121,481 (100)
Test method					
In-person	2,767 (23.2)	12,467 (25.7)^{m}	21,561 (44.7)	9,872 (77.1)^{m}	46,667 (38.4)
Telephone	8,134 (68.1)	34,496 (71.2)^{m}	25,028 (51.8)	1,221 (9.5)^{m}	68,879 (56.7)
NOS	1,045 (8.8)	1,498 (3.1)	1,686 (3.5)	1,706 (13.3)^{m}	5,935 (4.9)
Total	11,946 (100)	48,461 (100)	48,275 (100)	12,799 (100)	121,481 (100)
Wordlist					
List 1	1,950 (16.3)	6,286 (13)	5,029 (10.4)	1,135 (8.9)	14,400 (11.9)
List 2	1,413 (11.8)	5,386 (11.1)	4,189 (8.7)	349 (2.7)^{m}	11,337 (9.3)
List 3	1,363 (11.4)	5,207 (10.7)	4,033 (8.4)	181 (1.4)^{m}	10,784 (8.9)
List 4	1,319 (11)	5,925 (12.2)	4,273 (8.9)	342 (2.7)^{m}	11,859 (9.8)
List 5	1,139 (9.5)	5,602 (11.6)	4,163 (8.6)	221 (1.7)^{m}	11,125 (9.2)
List 6	2,445 (20.5)	11,008 (22.7)	12,578 (26.1)	4,036 (31.5)	30,067 (24.8)
List 7	1,167 (9.8)	4,582 (9.5)^{m}	7,401 (15.3)	3,444 (26.9)^{m}	16,594 (13.7)
List 8	1,150 (9.6)	4,465 (9.2)^{m}	6,609 (13.7)	3,091 (24.2)^{m}	15,315 (12.6)
Total	11,946 (100)	4,8461 (100)	48,275 (100)	12,799 (100)	121,481 (100)
Classification					
Impaired	37 (0.3)	1,033 (2.1)	2,258 (4.7)	2,643 (20.7)^{m}	5,971 (4.9)
Normal	11,909 (99.7)	47,428 (97.9)	46,017 (95.3)	10,156 (79.3)	115,510 (95.1)
Total	11,946 (100)	48,461 (100)	48,275 (100)	12,799 (100)	121,481 (100)

NOTE. The numbers of cases (N, %) are given for each age group and predictor variable (education level, race, gender, test method, and wordlist) as well as for classification result (normal or impaired). To look for categories of a predictor variable or age group with meaningful deviations from the expected number of cases (significant interactions), we computed effect size statistics (Cramer's phi, φ_c) for each cell of its χ^2 table (see Methods). Cells of the table with medium or large effect sizes are superscripted with the letters *m* and *l*, respectively.

Abbreviation: NOS, not otherwise specified.

Table 2
Characteristics of the clinical sample (group B) by FAST stage

Group B	Sample		MPI			Age			Years of education			Female	
FAST	N	%	μ	σ	g	μ	σ	g	μ	σ	g	%	Odds
1	94	21.3	60.4	11.6	—	69.1	12.1	—	15.1	3.6	—	34.5	—
2	70	15.9	52.5	15.0	0.60	71.1	12.3	−0.16	14.6	3.6	0.15	49.4	1.43
3	84	19.1	50.1	17.0	0.71	70.2	13.1	−0.08	15.0	4.1	0.02	49.4	1.76
4	125	28.3	36.7	16.0	1.66	74.0	12.3	−0.40	13.9	4.2	0.31	55.6	2.30
5	45	10.2	27.8	15.1	2.53	78.0	8.8	−0.79	12.0	5.6	0.72	54.3	1.64
6	23	5.2	18.4	9.8	3.70	78.9	7.3	−0.86	12.9	5.1	0.57	55.2	1.49
2–6	347	78.7	40.7	18.8	1.12	73.3	12.1	−0.35	14.0	4.4	0.26	53.4	1.55
All	441	100	44.9	19.2	—	72.4	12.2	—	14.2	4.3	—	52.4	—

NOTE. Sample size, MPI score, age, years of education, and female gender statistics are given for each FAST stage and are based on the first patient visit. All effect size statistics were computed by using FAST 1 (normal aging) as the reference group. Hedges' g effect size statistics were computed for MPI, age, and education; odds ratios were computed for female subjects (see Methods).

functional decline and dementia severity and over the full spectrum of AD [14,15]. FAST stages 1 to 7 can be approximately characterized as follows: 1, no subjective or objective functional impairment; 2, subjective functional impairment only; 3, objective functional impairment not meeting criteria for dementia; 4, mild dementia; 5, moderate dementia; 6, moderately severe dementia; and 7, severe dementia. Patients with FAST stage 7 are usually too impaired to perform cognitive assessment. The present analysis therefore characterized MPI score changes over FAST stages 1 to 6. Because MPI scores were not gaussian-distributed over each FAST stage (Fig. 4), we used a nonparametric K-sample test of the equality of medians (median, Stata 10.0) to test the null hypothesis of no significant differences among median MPI scores across FAST stages 1 to 6. The resulting probabilities that the median scores were equal were adjusted for multiple comparisons to test whether they were significant at a .05 level. Weighting these results by the reciprocal of the number of patient visits was also performed to examine the influence of repeated measures per patient.

2.11. Characterizing what the MPI score is measuring

Because there are substantive differences between the MPI score and total scores of the numbers of correctly recalled words during IFR and DFR, it is useful to characterize what the MPI score is measuring. First, we computed the proportion of MPI score variance explained by each of the CWL free recall total scores. IFR, DFR, and TFR scores were highly correlated with MPI score (R^2 = 0.7 to 0.84). However, because the MPI score substantially improves classification of normal versus MCI, the variance not explained by total scores contains useful information captured in the pattern of memory performance measured by the MPI.

The additional information captured by the MPI score can be visualized by applying correspondence analysis to the

Table 3
Effect sizes (Cohen's f^2) of predictor variables on MPI, IFR, DFR, and TFR scores (outcomes) for entire group A sample and for those classified as normal

Predictor	Predictor effect size[*] on outcome (by classification group)							
	MPI		IFR		DFR		TFR	
	All	NL	All	NL	All	NL	All	NL
Age group	0.682	0.671	0.073	0.044	0.077	0.044	0.085	0.051
Race[†]	0.003	0.004	0.004	0.005	0.005	0.005	0.005	0.006
Years of education	0.040	0.025	0.039	0.025	0.028	0.016	0.040	0.026
Gender	0.020	0.021	0.020	0.019	0.022	0.022	0.024	0.023
Test method[‡]	0.051	0.026	0.045	0.028	0.045	0.022	0.051	0.031
Wordlist	0.009	0.009	0.009	0.008	0.009	0.009	0.010	0.010
All but age	0.121	0.088	0.120	0.094	0.108	0.078	0.131	0.116
Variance explained by all predictors	55.0%	52.5%	24.9%	18.7%	23.4%	16.3%	26.9%	20.3%

NOTE. For the entire group A sample (N = 121,481) and for those subjects classified as normal (N = 115,510), the outcome measures of MPI score, CWL IFR, DFR, and TFR scores were each regressed against the full set of predictor variables (age, race, education, gender, test method, and wordlist). To determine each predictor's effect size, we removed it, re-ran the regression, and then computed Cohen's f^2 statistic. Effect size was also determined for all predictors but age combined. The percentage of variance explained by the full set of predictors was also determined for each outcome measure.

Abbreviation: NL, normal.

[*] The magnitude of the effect size of Cohen's f^2 statistic is negligible (<0.02), small (≥0.02 to <0.15), medium (≥0.15 to <0.35), and large (≥0.35).

[†] Group A sample size for race was <121,481 because of lack of requirement to report race. For normal + impaired, N = 8,033. For normal, N = 6,494.

[‡] Group A sample size for test method was 2,640 subjects per random sample to balance the number of subjects receiving the telephone versus in-person method of testing within each age group. Bootstrap random sampling with replacement by using 100 random samples was used to estimate test method effect size.

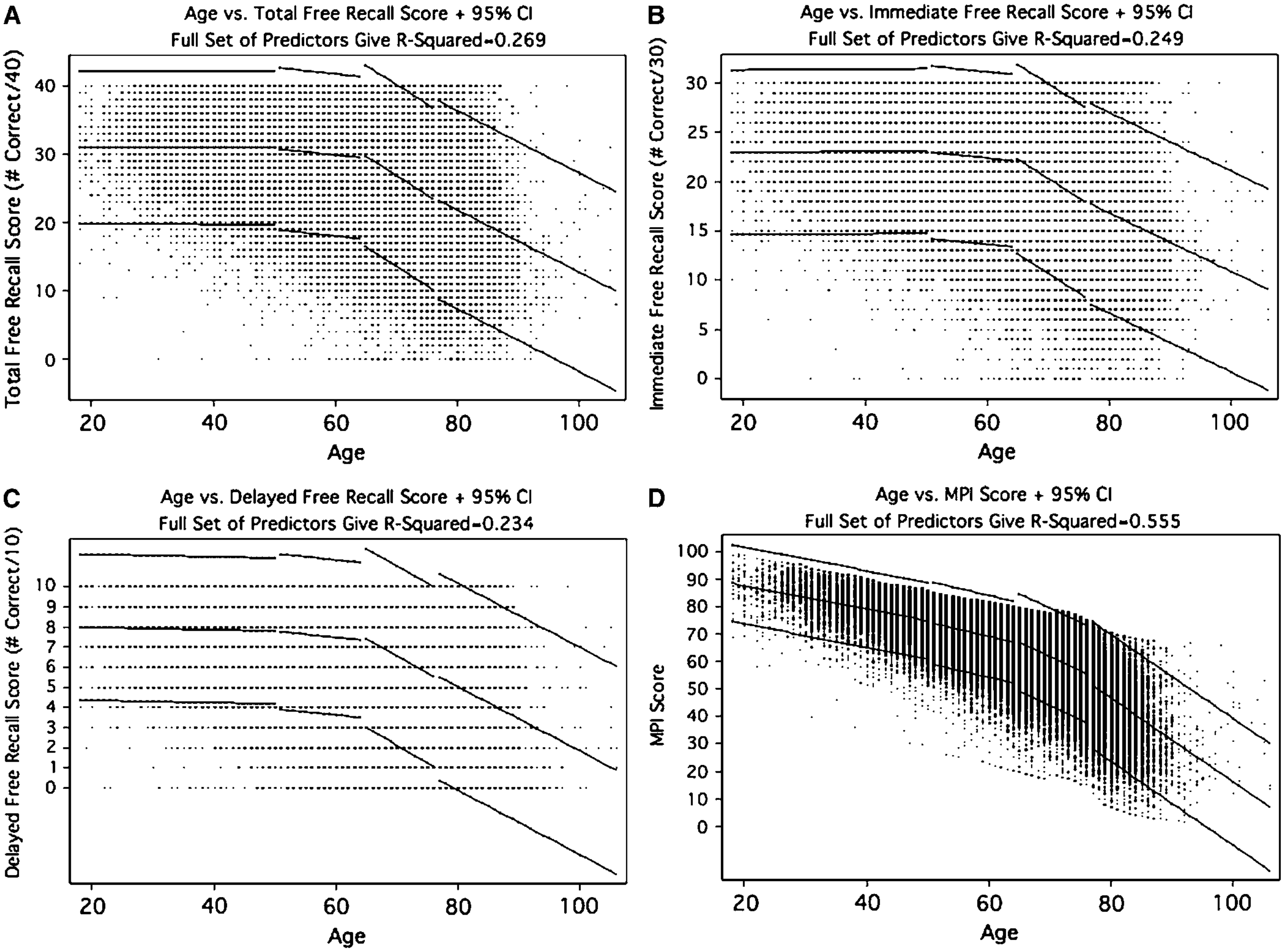

Fig. 3. Regressions of TFR (A), CWL IFR (B), DFR (C), and MPI (D) scores against the full set of predictors for the entire group A sample (N = 121,481).

entire group A sample. Correspondence analysis allows one to examine the effect of recalling or not recalling each word in each trial in relation to the entire sample. Each of these effects is represented as an optimal column score vector, and the effect of each case is represented as an optimal row score vector. We also included the age group categories as columns so that the effect of each age group could be examined. The first two dimensions (D1, D2) of these optimal column and row score vectors are plotted in Fig. 5 to show how age group and the recall or nonrecall of each word in each trial influences the classification of the cases as normal or impaired. There is one plot for each of the four trials. The cases are plotted as light gray dots for normal classification and as dark gray dots for impaired. The age groups are plotted as large dots (white, 18 to 50; light gray, 51 to 64; dark gray, 65 to 76; black, 77 to 106 years), and the recall or nonrecall of each word in a given trial is plotted as symbols labeled 1 to 10.

3. Results

In Table 1, the age group distributions of all predictor variables and of MCIS classification showed highly significant deviations from their expected frequencies across the four age groups. Examination of Cramer's φ_c statistic for medium (0.15 to 0.35) or large (>0.35) effect sizes showed the following:

1. Education: no medium or large effects of education;
2. Race: a medium effect for whites aged 77 to 106 years;
3. Test method: medium effects for both in-person and telephone administration for age groups 51 to 64 and 77 to 106 years;
4. Wordlist: medium effect for wordlists 1 to 5, 7, and 8 for age group 77 to 106 years and medium effect for wordlists 7 and 8 for age group 51 to 64 years;
5. MCIS classification: a medium effect of class, Impaired, for age group 77 to 106 years.

In Table 2, the 441 subjects in group B had no statistically significant differences in either education or gender across FAST stages. There were statistically significant age differences ($P < .05$, Bonferroni adjusted) between FAST stages 1 and 4 to 6, between stages 2 and 5 to 6, and between stages 3 and 6. Median MPI scores significantly differed between all FAST stages ($P < .0002$, Bonferroni adjusted).

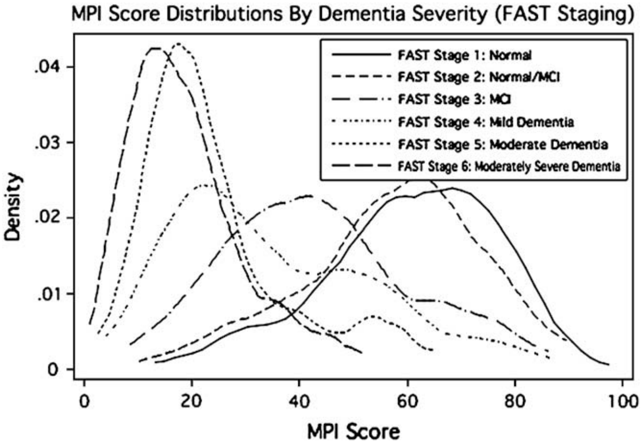

Fig. 4. MPI score distributions of FAST stages 1–6 for the entire group B clinical sample (N = 441 patients with 2,416 assessments).

In terms of the effect of age, Fig. 5 shows that the older the age group, the greater the likelihood of being impaired (older age groups are to the left of younger age groups). In terms of the effect of recalling a given word, symbols labeled 1 to 10 on the right side of each graph show that words 1, 9, and 10 have approximately the same D1 value, but word 1 has a higher D2 value than words 9 and 10. Words 2 to 6 have approximately the same D1 value but different D2 values. The effect of repeated learning in Trials 2 and 3 (the second and third graphs) on these optimal scores is to bring them closer together to form three distinct clusters: word 1 alone, words 7 to 10 together, and words 2 to 6 together. The effect of the delay on recall in Trial 4 (the last graph) is to give words 2 to 10 very similar D1 values but different D2 values. Word 1 continues to remain separate (easiest to recall) from the other list words.

The optimal scores for not recalling the 10 words are the symbols labeled 1 to 10 on the left side of each graph. Not recalling word 1 always has a greater value associated with it than the other words, regardless of the trial in which it occurs. Also, the words are approximately ordered 1 to 10 from highest to lowest along the D2 axis, and the ordering becomes increasingly sequential with repeated learning (IFR Trials 1 to 3). After a delay (DFR Trial 4), the ordering of the nonrecalled words becomes slightly less sequential.

Figure 6 shows that both CWL TFR and DFR data and MPI score data distributions fit reasonably closely to a gaussian distribution, making them amenable to standard parametric statistical procedures.

Figure 3 shows the age distributions of the outcome variables MPI score, CWL TFR, IFR, and DFR scores, plotted against age for the entire group A sample. To visualize outcome measure differences in explanatory power, the black lines in each graph are the age-dependent mean scores and their 95% CIs as determined by piecewise linear regression of each of the four age groups against each outcome measure. To a first approximation, this visualization is correct because the variance explained by all other predictors was similar for the four outcome measures.

The R^2 values reported in the title of each graph are based on regression of the outcome variable against the full set of predictors. The three CWL free recall graphs show that given a particular free recall score, one cannot accurately predict a person's age, and that given a person's age, one cannot accurately predict their free recall score. In contrast, the MPI graph shows that either of these predictions can be much improved.

Table 3 shows the effect sizes of the predictor variables of MPI and CWL IFR, DFR, and TFR for the entire group A sample as well as for those classified as normal. The last row of the table shows the percentage of variance explained by the full set of predictor variables. In terms of the amount of variance explained by the predictor variables, they accounted for 2 to 3 times more MPI score variance than for the variance of the free recall score variables. In terms of predictor effect sizes, the magnitude of their Cohen's f^2 statistics was classified by using the following qualitative scale [10]: negligible (<0.02), small (≥0.02 to <0.15), medium (≥0.15 to <0.35), and large (≥0.35). The effect size of age on MPI score was large, and it was small on IFR, DFR, and TFR scores. Predictors with negligible effect sizes on all outcome score measures were race and wordlist. Predictors with small effect sizes were education, gender, and test method. The effect size of all predictors minus age was small.

Figure 4 shows the MPI score distributions for each of the FAST stages 1 to 6. Visual inspection of these graphs suggests that they are not normally distributed and justifies the use of a nonparametric method to test for significant MPI score differences between them.

4. Discussion

The primary findings of the present article are the following:

1. MPI score accounts for 2 to 3 times more variance than currently used scoring methods.
2. MPI score decreases with increasing dementia severity as measured by FAST staging.
3. Age has a large effect on MPI score prediction but has only a small effect on predicting currently used scoring methods (ie, IFR, DFR, TFR). This suggests that scores based on numbers of words recalled are less informative than scores based on recall patterns.
4. The combined effect size of other predictors (gender, education, race, test method, and wordlist) is small for all outcome measures studied.
5. Recalling versus not recalling list-words across CWL trials contributes differently toward classifying normal and impaired persons.
   a. Recalling the first list-word has an effect that is distinct from all other words and is analogous to a primacy effect.

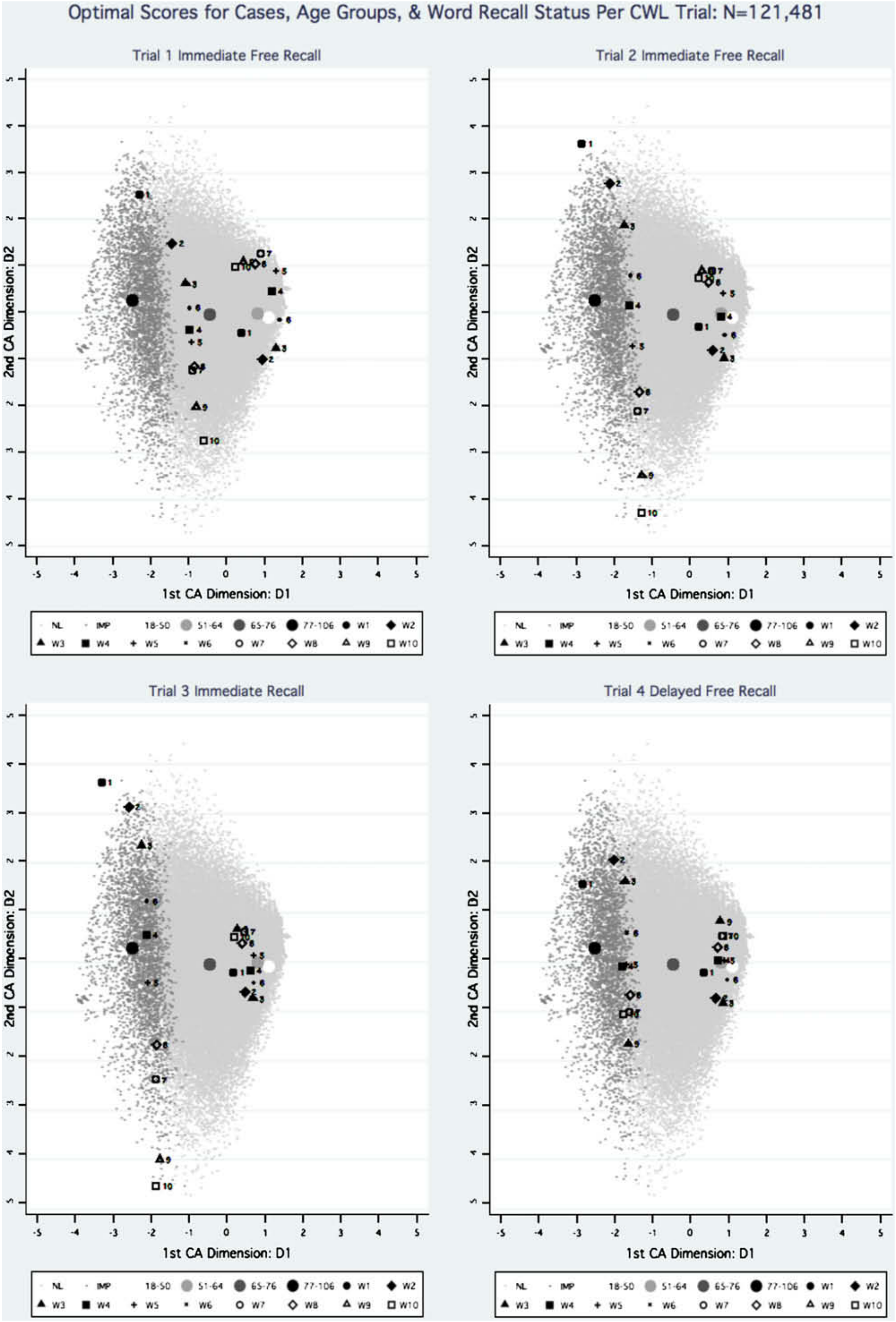

Fig. 5. Optimal scores (dimensions 1 and 2) for cases, age groups, and item responses (recalled or not recalled) for the 10 list-words in each CWL trial. The four plots are for CWL Trials 1 to 4. The first two ortho-normal dimensions of the optimal scores for the normal (light gray dots) and impaired (dark gray dots) cases are plotted as background to highlight their relationship to the optimal scores for each age group and for each of the 10 words recalled (the symbols labeled 1 to 10 on the right side of the plot) or not recalled (the symbols labeled 1 to 10 on the left side) in each CWL trial. Optimal scores lying close to each other are more similar in the information they provide for predicting normal and impaired cases.

b. Not recalling the first list-word also has an effect that is distinct from all other words.
c. Repeated learning of the CWL causes words recalled from the middle of the list to cluster together and separately clusters words recalled from the end of the list.
d. The effect of a delay of several minutes before the DFR task is to spread apart the words within these clusters.
e. The effect of not recalling list-words is not simply the opposite of recalling them. The optimal scores

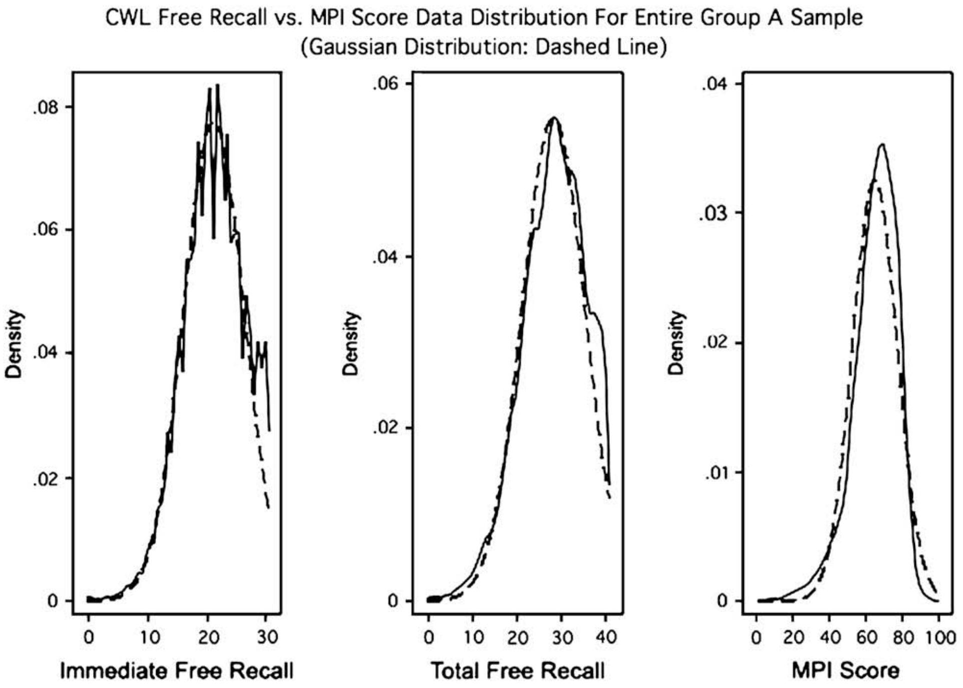

Fig. 6. Probability distributions of CWL IFR and TFR scores and MPI scores for initial assessments of the entire group A sample (N = 121,481) (gaussian distribution, dashed line).

of nonrecalled words follow closely a temporally ordered sequence along the second optimal score dimension, with word 1 having the highest value and word 10 the lowest value.

The implication of the MPI score decreasing with increasing dementia severity is that if a person's MPI score declines over time, it signifies increasing severity of dementia or cognitively related functional impairment. To increase classification accuracy further, persons whose MPI scores decline over time but are still in the normal range need to be subclassified into those whose decline is due to aging versus those whose decline is due to a dementing disorder. Such classification is complex and will be the subject of a future analysis with more sensitive biomarker indices of asymptomatic disease states. To distinguish age-related versus disease-related change will require combining the methodology of graphical hierarchical bayesian methods with cognitive processing models and then applying them to multiple data sets with the necessary attributes.

The finding that for the same test data, the MPI score explains 2 to 3 times more variance than the CWL recall score measures is important in terms of measuring treatment effect and increasing classification accuracy. The greater reduction in the variability of the MPI score is accomplished through greater explanatory power of the covariates (predictor variables), which explains ∼55% of the MPI score's variance but explains less than 27% of the variance of the CWL total score measures. The fact that Fig. 5 shows different effects of recalling or not recalling each word in each trial means that ignoring this information (which is what the total score measures do) increases the unexplained variance and lowers classification accuracy and measurement of treatment effect.

Why is the pattern of CERAD wordlist performance more informative than total number of words recalled after a delay, which has been the standard way of scoring memory tests? There are several potential answers derived from different perspectives. From a pure information theoretical perspective, the number of ways of recalling 10 words across four trials is 5.9×10^{47}, whereas the number of possible total scores on the delayed recall trial is 11. This means that there is much information being thrown away by reliance on a total score, and that some of this information is relevant to the task of classification and measurement. From a mathematical psychological perspective, the MPI scoring algorithm accounts for differential scoring of an item according to its position in the wordlist, to the number of times it has been presented, to the presence of a delay before recall, and to whether it was recalled. In our original study [5], this differential scoring of an item improved the ability to discriminate normal aging from MCI by approximately 12% when compared with the DFR score. This means that these attributes of recall performance are important in discriminating a disease process affecting memory from that caused by normal aging. In the mathematical psychology literature, computational models of the order of recall have shown that these attributes improve the model's predictive ability [8].

Currently, the most commonly used tests of memory (eg, California Verbal Learning Test, Hopkins Verbal Learning Test, Rey Auditory Verbal Learning Test, Wechsler Logical

Memory Scale, cognitive portion of the Alzheimer's Disease Assessment Scale, standard scoring of the CWL) use the number of words freely recalled, with or without adjustments for age, gender, and education, to interpret results. The present study suggests that it is worth examining the pattern of recalled and not recalled items in other memory tests to see whether the proportion of explained variance and classification accuracy can be further improved. It would also be worth analyzing the wordlists used by these tests to determine whether their effect sizes on the outcomes are negligible.

Explaining more of the residual variance in free recall tasks among normal as well as cognitively impaired or demented individuals can do the following:

1. Shrink the confidence band around expected rates of change for a given subject group.
2. Help determine whether the expected change in a subject's scores are:
   a. Within the expected confidence band, suggesting the subject is still normal.
   b. Significantly greater than the upper bound of the confidence band, suggesting a treatment effect.
   c. Significantly below the lower bound of the confidence band, suggesting a dementing disorder.

To produce such confidence bands for normal aging-related changes in memory is a nontrivial issue that involves careful analysis of various sources of measurement error and will be addressed in a separate study with the hierarchical bayesian methods previously described.

Age, race, gender, education, test method, and wordlist, either alone or combined, were negligible to small predictors (negligible to small effect sizes) of the variation in CWL scores of IFR, DFR, and TFR. In contrast, the effect size of age on MPI score was large, and MPI variance explained by the full set of predictors was 2 to 3 times greater than that obtained for the CWL free recall scores. These findings are consistent with the MPI score having more of its total variance explained (a higher signal:noise ratio) than do the CWL free recall scores.

The improved measurement characteristics of the MPI score over other scores based on the numbers of words correctly recalled can facilitate more accurate detection of the transition from normal aging to MCI and help distinguish treatment and disease progression effects from those caused by normal aging.

Acknowledgments

We greatly appreciate Dr Doug Trenkle for sharing the dataset from the Hancock County Aging Project and carefully reviewing the manuscript.

All authors are employees of Medical Care Corporation. Drs Shankle and Hara are the trustees of the trust that holds shares of Medical Care Corporation.

References

[1] Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): part I—clinical and neuropsychological assessment of Alzheimer's disease. Neurology 1989;39:1159–65.

[2] Lamberty GJ, Kennedy CM, Flashman LA. Clinical utility of the CERAD word list memory test. Appl Neuropsychol 1995;2:170–3.

[3] Andel R, McCleary CA, Murdock GA, Fiske A, Wilcox RR, Gatz M. Performance on the CERAD Word List Memory task: a comparison of university-based and community-based groups. Int J Geriatr Psychiatry 2003;18:733–9.

[4] Karrasch M, Sinerva E, Gronholm P, Rinne J, Laine M. CERAD test performances in amnestic mild cognitive impairment and Alzheimer's disease. Acta Neurol Scand 2005;111:172–9.

[5] Shankle WR, Romney AK, Hara J, Fortier D, Dick M, Chen J, et al. Method to improve the detection of mild cognitive impairment. Proc Natl Acad Sci USA 2005;102:4919–4.

[6] Trenkle D, Shankle WR, Azen SP. Detecting cognitive impairment in primary care: performance assessment of three screening instruments. J Alzheimers Dis 2007;11:323–35.

[7] Cho A, Sugimura M, Nakano S, Yamada T. Early detection and diagnosis of MCI using the MCI screen test. Jpn J Clin Exp Med 2007;84:1152–60.

[8] Howard MW, Kahana MJ. A distributed representation of temporal context. J Mathematical Psychology 2002;46:269–99.

[9] Weller SC, Romney AK. Metric scaling: correspondence analysis. Newbury Park, CA: SAGE Publications; 1990.

[10] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.

[11] Kukull WA, Larson EB, Reifler BV, Lampe TH, Yerby MS, Hughes JP. The validity of 3 clinical diagnostic criteria for Alzheimer's disease. Neurology 1990;40:1364–9.

[12] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 2007; 6:667–9.

[13] Reisberg B, Saeed MU. Comprehensive textbook of geriatric psychiatry. 3rd ed. New York, NY: WW Norton & Co; 2004.

[14] Reisberg B. Functional assessment staging (FAST). Psychopharmacol Bull 1988;24:653–9.

[15] Sclan SG, Reisberg B. Functional assessment staging (FAST) in Alzheimer's disease: reliability, validity, and ordinality. Int Psychogeriatr 1992;4(Suppl 1):55–69.

[16] Hedges LV, Olkin I. Statistical methods for meta-analysis. San Diego, CA: Academic Press; 1985.