Published in final edited form as:

J Int Neuropsychol Soc. 2011 July ; 17(4): 625–638. doi:10.1017/S1355617711000476.

Explaining Differences in Episodic Memory Performance among Older African Americans and Whites: The Roles of Factors Related to Cognitive Reserve and Test Bias

Denise C. Fyffe¹, Shubhabrata Mukherjee², Lisa L. Barnes³, Jennifer J. Manly⁴, David A. Bennett³, and Paul K. Crane²

¹Spinal Cord Injury/Outcomes & Assessment Research Laboratory, Kessler Foundation Research Center, West Orange, New Jersey and Physical Medicine and Rehabilitation, New Jersey Medical School University of Medicine and Dentistry of New Jersey, Newark, New Jersey

²School of Medicine, University of Washington, Seattle, Washington

³Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois

⁴Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, New York, New York

Abstract

Older African Americans tend to perform poorly in comparison with older Whites on episodic memory tests. Observed group differences may reflect some combination of biological differences, measurement bias, and other confounding factors that differ across groups. Cognitive reserve refers to the hypothesis that factors, such as years of education, cognitive activity, and socioeconomic status, promote brain resilience in the face of pathological threats to brain integrity in late life. Educational quality, measured by reading test performance, has been postulated as an important aspect of cognitive reserve. Previous studies have not concurrently evaluated test bias and other explanations for observed differences between older African Americans and Whites. We combined data from two studies to address this question. We analyzed data from 273 African American and 720 White older adults. We assessed DIF using an item response theory/ordinal logistic regression approach. DIF and factors associated with cognitive reserve did not explain the relationship between race, and age- and sex-adjusted episodic memory test performance. However, reading level did explain this relationship. The results reinforce the importance of considering education quality, as measured by reading level, when assessing cognition among diverse older adults.

Keywords

Mental recall; Ethnic groups; Psychometrics; Cognition; Education; Health status disparities

INTRODUCTION

Racially disparate outcomes on neuropsychological episodic memory tests have persistently been observed among older adults. Generally, older African Americans demonstrate lower

scores on episodic memory tests than Whites (Fillenbaum, Peterson, Welsh-Bohmer, Kukull, & Heyman, 1998; Manly et al., 1998; Masel & Peek, 2009; McDougall, Vaughan, Acee, & Becker, 2007; Schwartz et al., 2004; Whitfield et al., 2000; Zsembik & Peek, 2001). Worse performance may represent poorer episodic memory functioning, measurement problems such as test bias, or a combination. Poor performance among African Americans due to measurement problems could lead to misdiagnosis of memory disorders (Gurland et al., 1999; Weiner, 2008; Whitfield, 2002; Whitfield et al., 2000). Inaccurate assessment and inappropriate diagnoses can have profound negative implications on quality of life, end of life decision making, and caregiver support (Dilworth-Anderson, Hendrie, Manly, Khachaturian, & Fazio, 2008; Parker & Philp, 2004). Previous investigators have identified demographic characteristics including age and sex (Manly et al., 1998; McDougall, et al., 2007; Mungas, Reed, Farias, & DeCarli, 2009; Zsembik & Peek, 2001), health conditions including hypertension and cardiovascular disease (Schwartz et al., 2004; Whitfield et al., 2000), and sociocultural variables including education, language, acculturation, and socioeconomic status (Boone, Victor, Wen, Razani, & Ponton, 2007; Manly, Byrd, Touradji, & Stern, 2004) as factors associated with observed score differences across groups.

Stern et al., suggested educational experiences influence brain development and can be considered a proxy for cognitive reserve (Stern et al., 1994; Stern, 2009). Parental education (Kaplan et al., 2001; Rogers et al., 2009; Singh-Manoux, Richards, & Marmot, 2005), home experiences that stimulate childhood learning (Everson-Rose, Mendes de Leon, Bienias, Wilson, & Evans, 2003), and lifetime engagement in cognitive activities (Scarmeas & Stern, 2003; Wilson, Barnes, & Bennett, 2003; Wilson et al., 2005) are examples of factors found to influence late-life cognitive functioning. These experiences, conceptualized as cognitive reserve in the current manuscript, may preserve cognitive functioning in the face of brain pathology in later life (Jones et al., 2010; Scarmeas & Stern, 2003). The primary goal of this study is to examine factors associated with cognitive reserve concurrently for measurement bias and their ability to explain differences in episodic memory performance across African Americans and Whites.

The association between education and reserve may be partially mediated by socioeconomic status and education quality (Brunner, 2005; Dotson, Kitner-Triolo, Evans, & Zonderman, 2009; Kaplan et al., 2001; Stern, Albert, Tang, & Tsai, 1999). Higher socioeconomic status may afford opportunities to engage in cognitively stimulating experiences, which may buffer against late life cognitive decline (Stern et al., 1994, 1999; Stern, 2006). Manly, Touradji, Tang, and Stern (2003) and Manly, Schupf, Tang, and Stern (2005) studied education quality as measured by performance on reading tests (Cosentino, Manly, & Mungas, 2007). Low reading levels (i.e., a proxy for poor education quality) were associated with more rapid rates of cognitive decline (Manly et al., 2003, 2005).

Demographic, health, and sociocultural factors that contribute to differential episodic memory ability may represent test bias (Brickman, Cabo, & Manly, 2006; Gasquoine, 2009; Pedraza & Mungas, 2008; Robertson, Liner, & Heaton, 2009; Rosselli & Ardila, 2003). Educational experiences that lead to the acquisition of test-taking strategies can increase “test wiseness” and may inflate test scores (Gasquoine, 2009; Manly, Jacobs, Touradji, Small, & Stern, 2002; Robertson et al., 2009; Rosselli & Ardila, 2003; Scruggs & Lifson, 1985). If test wiseness varies across groups, individuals in different groups with the same underlying level of the ability measured by the test would have unequal expected scores, which is a definition of differential item functioning (DIF) (Camilli & Shepard, 1994; Thissen, Steinberg, & Wainer, 1993). Other factors representing test bias include reaction to test content (e.g., familiarity, interest) (Brickman et al., 2006; Flaughner, 1978; Stricker & Emmerich, 1999; Teng & Manly, 2005) and cultural factors including stereotype threat,

language, or unrepresentative norms (Brickman et al., 2006; Gasquoine, 2009; Kit, Tuokko, & Mateer, 2008; Loewenstein, Arguelles, Arguelles, & Linn-Fuentes, 1994; Manly et al., 2002; Manly, 2008; Teng & Manly, 2005; Whitfield, 2002).

Meaningful comparisons of performance across groups necessitate attention to measurement equivalence (Teresi, Kleinman, & Ocepek-Welikson, 2000; Teresi, Stewart, Morales, & Stahl, 2006; Tuokko et al., 2009). Several researchers have applied DIF methodology to assess relationships between characteristics associated with test bias and performance on neuropsychological tests among racially diverse older adults (Crane, van Belle, & Larson, 2004; Crane et al., 2008; Jones, 2003; Pedraza et al., 2009; Ramirez, Teresi, Holmes, Gurland, & Lantigua, 2006; Teresi, Holmes, Ramirez, Gurland, & Lantigua, 2001; Teresi et al., 1995). Much of this previous work has found substantial DIF in global measures of cognition, such as the Mini-Mental State Examination (MMSE) (Crane, Gibbons, Jolley, & van Belle, 2006; Dorans & Kulick, 2006; Jones, 2006; Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006; Ramirez et al., 2006) or the Cognitive Abilities Screening Instrument (CASI) (Crane et al., 2004; Gibbons et al., 2009). DIF has also been observed in specific cognitive domains, such as visual naming ability (Pedraza et al., 2009), fluency, and working memory (Crane et al., 2008). To our knowledge this is the first study to examine DIF in African Americans and Whites on a measure of episodic memory.

DIF analyses determine whether individual characteristics exaggerate or attenuate the probability of successful responses to episodic memory items, given a particular level of episodic memory functioning. DIF analyses often focus on item-level findings. Crane, Gibbons, Narasimhalu, Lai and Cella (2007) and Crane, Gibbons, and Ocepek-Welikson, et al. (2007) suggest there may be different audiences for DIF analyses. Scale developers may be most interested in item-level findings. Clinicians may be primarily interested in individual-level DIF impact. Social scientists may be primarily interested in group-level DIF impact, which addresses the question, “Is it likely that DIF might impact mean scores for groups or relationships between covariates of interest?” (Crane, Gibbons, and Ocepek-Welikson, et al., 2007; Crane, Gibbons, Narasimhalu, et al., 2007). In this study, we are primarily interested in group-level DIF impact. One research question being posed is: Does DIF impact the relationships between factors associated with reserve and episodic memory functioning across African American and White older adults?

Figure 1 depicts theorized relationships evaluated in this study. Observed variables (performance on episodic memory tests, demographics, indicators associated with reserve) are depicted in rectangles, while the unobserved factor (actual episodic memory functioning) is in an oval. The prior work of Manly et al. (2002, 2003, 2005) suggested that educational experiences were particularly important. Because these investigators did not test for DIF, its possible importance as an explanatory factor is unknown. In the current study we directly tested for DIF and depict DIF in a dashed box in Figure 1. The dashed box indicates that usually DIF is ignored, but is included in the present study. Thus, the goals of this study are thus to better understand relationships between memory performance and demographic and cognitive reserve covariates, while accounting for DIF.

METHOD

Participants

Study participants were identified from the Memory and Aging Project (MAP) and the Minority Aging Research Study (MARS) conducted by the Rush Alzheimer’s Disease Center. MAP and MARS are ongoing longitudinal cohort studies among community-dwelling older adults in Chicago. MAP began enrollment in 1997 (Bennett et al., 2005). Consenting participants agreed to detailed annual evaluations, cognitive testing, and

postmortem organ donation. MARS has a nearly identical design and began enrollment of African Americans in 2004. By April 2010, MAP included 1304 participants, and MARS 349. Recruitment strategies were so similar that a few African Americans are enrolled in both studies.

We evaluated baseline data from self-identified African Americans or Whites who were free of dementia, and had complete episodic memory and cognitive reserve data. The data from these studies were obtained in compliance with Rush's Institutional Review Board regulations.

Clinical Evaluations

Participants completed clinical evaluations including medical history, neurological examination, and neuropsychological assessment (Arvanitakis, Bennett, Wilson, & Barnes, 2010; Bennett et al., 2005). A clinician used clinical data and standard criteria to classify dementia and Alzheimer's Disease (McKhann et al., 1984).

Neuropsychological Evaluations

Participants completed a 19-test battery assessing five cognitive domains. We evaluated episodic memory tests common across MAP and MARS. (a) Story recall (4 scores). Logical Memory Story A (Wechsler, 1987) is a fact-dense textual passage read aloud once; the participant is asked to recall elements immediately and after a delay. The East Boston Memory Test (Albert et al., 1991) is similar, and includes scores for immediate and delayed recall. (b) Word list (3 scores). The 10-word CERAD list (Morris et al., 1989) was administered in three learning trials that are summed (range, 0–30). After a distracter task, the participant is asked to recall the words (range, 0–10). Participants are then presented with ten trials of four words, and asked to identify the one on the CERAD list (range, 0–10).

Cognitive reserve—Cognitive reserve indicators included: years of personal, maternal, and paternal education, childhood cognitive activity frequency, income at age 40, and education quality, as measured by reading level (see below). We initially categorized self-reported personal years of education as (1) some primary (<grade 8); (2) primary (completed grade 8); (3) some high school (9–11); (4) high school (completed grade 12); or (5) post-secondary (13 or greater). For DIF analyses, we categorized education as <12 and 12 years to ensure adequate analytic sample sizes.

We calculated childhood cognitive activity from self-reported activities at ages 6 and 12. Participants were asked how often someone read to them, told them stories, or played games with them (age 6) and how often they read books and magazines or went to the library (age 12); response options ranged from less than once a year (1 point) to almost every day (5 points) and composite scores were obtained by averaging across the five items (Wilson et al., 2003). The scale has demonstrated adequate psychometric properties (Cronbach's $\alpha=0.88$; test-retest reliability of $r=0.79$) in studies with older adults (Barnes, Wilson, de Leon, & Bennett, 2006; Wilson et al., 2005). We dichotomized average scores at 3 and >3 activities to ensure adequate analytic sample sizes.

Income at age 40 was reported in one of six categories defined by a range of dollar amounts. We compared participant responses to the median U.S. family income for the appropriate year (United States Census Bureau, 2010). We categorized income as below or above median income at age 40.

Reading level was measured by reading tests. MAP participants were administered the National Adult Reading Test (NART) (Nelson, 1982), while MARS participants were

administered the third edition of the Wide Range Achievement Test Reading subtest (WRAT-3) (Wilkinson, 1993). For each test, participants read aloud words of increasing complexity; correct pronunciation is required to obtain a point.

We analyzed NART and WRAT-3 data from the 10 individuals enrolled in both studies to co-calibrate this variable. We identified 23 data points where those individuals were evaluated by the two tests at least two times within a 6-month window. For those 23 occasions, we examined a scatterplot (Appendix 1) that confirmed *Z* scores on the two tests appeared to be roughly linearly related to each other. We identified the median *Z* score on the WRAT-3 and the median *Z* score on the NART for these individuals, and used those *Z* scores to categorize reading levels from the parent studies.

Data Analysis

Overview—We derived three different composite scores from the seven episodic memory test data points: a composite *Z*-score, an IRT score that ignored DIF (a “naive” score), and an IRT score that accounted for DIF with respect to all of the covariates. We performed linear regression analyses using standardized composite scores as dependent variables and race as the primary predictor. We included demographic factors, and factors associated with cognitive reserve, paying particular attention to reading level. We performed a series of sensitivity analyses to assess the robustness of our findings.

Composite *Z* score—We created the composite measure of episodic memory by converting raw scores on each test to *Z* scores using the baseline MAP mean and standard deviation. We averaged these *Z* scores (Wilson et al., 2003, 2005).

Dimensionality

Both the naive IRT score and the IRT score accounting for multiple sources of DIF rely on an assumption of unidimensionality, that is, that the items can be conceptualized as measuring a single underlying construct. There is no single standard approach for determining whether a scale is sufficiently unidimensional. We used exploratory and confirmatory factor analyses.

Naive IRT scores—We used Parscale (Muraki & Bock, 2003) using Samejima’s graded response model (Samejima, 1969) and *expected a posteriori* (EAP) scoring. The graded response model is a polytomous extension of the two-parameter logistic model (2PL) (Lord & Novick, 1968).

IRT scores that accounted for all forms of DIF—We used a hybrid ordinal logistic regression/IRT approach to identify and account for DIF, using *difwithpar* software (Crane et al., 2006). We analyzed several covariates for DIF: self-reported race, sex, education, age, father’s education, mother’s education, childhood cognitive activities, income at age 40, and reading level. We were primarily interested in accounting for all sources of DIF. Detailed methods have been published previously (Crane et al., 2006, 2008).

Regression analyses—All regression models included an indicator term for race. We transformed each episodic memory composite score to have a mean of 0 and standard deviation of 1. We performed a series of regression analyses with the composite episodic memory scores as dependent variables: (1) Base: race; (2) Demographics: race plus demographics (sex and age); (3) Demographics and cognitive reserve except reading level: model 2 plus cognitive reserve factors other than reading level (years of education, father’s education, mother’s education, childhood cognitive activities, and income at age 40); (4) Demographics and cognitive reserve including reading level: model 3 plus reading level.

Sensitivity analyses—We performed several sensitivity analyses to determine whether assumptions made in our modeling affected our conclusions. We repeated DIF analyses related to race using Multiple Indicator Multiple Cause (MIMIC) modeling. These analyses were performed in two ways, using (1) a single factor model (analogous to the IRT approach used in the primary analysis); and, (2) a bi-factor model that does not rely on the assumption of unidimensionality.

We assessed multicollinearity between the covariates. We matched African Americans to Whites of similar age and education and the same sex, and repeated the regression analyses to control for cohort effects. We performed regression analyses with age, education, and childhood cognitive activity as continuous variables. The scores we used to co-calibrate the reading tests may lead to misclassifying high or low reading levels (Appendix 1), so we performed a secondary analysis in which we omitted people whose reading scores were close to the cutoff values (within 0.25 *SD* of the cutoff values), that is, people whose reading levels were most likely to be misclassified to ensure that misclassification of reading level was not driving the results.

We performed additional analyses to determine whether the reading level effect was unique, or whether using another cognitive test would have the same effect. We compared correlations between reading scores and Digit Span Forward, Digit Span Backward (Wechsler Memory Test-R) (Wechsler, 1987), and Digit Ordering (Cooper & Sagar, 1993; Wilson et al., 2002). We used Digit Ordering, the test that had the lowest correlations with reading scores, to avoid confounding the domains. We dichotomized Digit Ordering so similar proportions would be classified as high or low as were in those categories for reading level. We then repeated the final regression model replacing reading level with Digit Ordering.

RESULTS

Demographics and Episodic Memory Scores

Data were available from 1644 participants. We performed our primary analyses on the 993 participants with complete data, including 273 African Americans and 720 Whites. Some participants who were included in the data set also self-identified as Hispanic: 5 (2%) of the African Americans and 77 (11%) of the Whites. Figure 2 provides an outline of the sample derivation. There were 83 participants excluded due to a diagnosis of Alzheimer's disease or other dementia and 12 participants excluded because they self-identified in a racial group other than African American or White. An additional 556 participants were excluded because they had missing data. Missing data were especially prevalent for three reserve indicators: mother's and father's education and income at 40. The demographic and episodic memory characteristics remained the same when we included participants with missing data. We also compared results from the 993 people with complete data on all covariates to results from the 1421 people with data on all covariates other than mother's education, father's education, and income at age 40, and all regression coefficients were within a few hundredths of each other.

The 993 participants in our primary analyses had a mean age of 77.8 years (*SD* = 7.6) and a mean of 14.8 years of education (*SD* = 3.3); 71% were women and 73% were White. Further demographic details are provided in Table 1. On average African Americans were younger and had more years of formal schooling than Whites, had approximately the same levels of parental education and income at age 40, had higher childhood cognitive activity scores, and had lower reading levels. Mean scores for African Americans and Whites for the individual episodic memory tests are shown in Table 1. The tests used to make episodic memory scores

demonstrated adequate reliability ($\alpha = 0.81$) and bivariate correlations ranging from 0.23 to 0.85.

IRT and DIF Analyses

We calculated three composite episodic memory scores, which were highly correlated. The two IRT scores were more closely correlated with each other ($r = 0.998$) than with the composite Z score ($r = 0.913$ for the naive IRT score and $r = 0.900$ for the IRT score accounting for DIF). Results from exploratory and confirmatory factor analyses indicated that the episodic memory indicators were sufficiently unidimensional for use of IRT. Only a single Eigen value was above 1 and the second factor had a negligible Eigen value, a single factor model did not fit well, so we fit a bi-factor model in which the three word list items formed a secondary factor and in which we allowed for residual correlation between the two Logical Memory items and similarly for the two East Boston items. This model fit well. Factor loadings between the single factor model and the bi-factor model were very similar, and all of the loadings on the general factor in the bi-factor model were >0.30 , which McDonald suggests is evidence of sufficient unidimensionality (McDonald, 1999).

The DIF analyses considered nine covariates: race, age, sex, education, income at age 40, early life cognitive activities, mother's education, father's education, and reading level. The difference between the IRT score accounting for all nine sources of DIF and the naive IRT score represents individual-level DIF impact. When DIF has a negligible impact, the difference will be close to zero. If DIF makes a big impact, this difference will be large. We compared differences to the median standard error of measurement for IRT scores in this data set, which was 0.3. Accounting for all sources of DIF led to changes larger than 0.3 for only six participants ($<1\%$), which suggested the overall individual level DIF impact was negligible.

We compared scores for African Americans and Whites when accounting for and ignoring DIF. The mean (SD) naive score for African Americans was -0.005 (0.88), and for Whites it was $+0.002$ (1.04), a difference of 0.007. The mean (SD) scores accounting for DIF for African Americans was -0.036 (0.89), and for Whites it was 0.014 (1.04), a difference of 0.050. Ignoring DIF thus very modestly attenuated differences in mean episodic memory scores between Whites and African Americans.

Factors Associated With Episodic Memory Scores

Regression results are shown in Table 2. The cells show values for regression coefficients for each model. The four sections show results obtained from models with: (1) race only; (2) race and demographics; (3) race, demographics and measures of cognitive reserve except reading level; and (4) race, demographics, and all measures of cognitive reserve including reading level. The three columns show results for the three different dependent variables (naive IRT score, IRT score accounting for all sources of DIF, and composite Z -score) used for the regression models.

Our primary focus in these analyses was on the coefficients associated with race, shown in the top row of each section of Table 2. The intercept term provides an estimate of the adjusted mean for the reference group, while the coefficient for race provides an estimate of the adjusted mean difference between African Americans and Whites. In unadjusted models, mean episodic memory scores were not different across race groups in our sample (Model 1). When we accounted for demographic differences across race groups by including age and sex, African Americans on average did worse than Whites (Model 2). These findings were consistent across the three composite episodic memory scores. We entered age and sex separately in the models and confirmed our suspicion that this effect was attributable to age.

The third section in Table 2 summarizes regression findings from models that included race, demographics, and measures of cognitive reserve other than reading level (Model 3). The coefficient for race was not affected by including these factors in the model, suggesting that differences across racial groups in age- and sex-adjusted episodic memory performance were not due to these factors. Again, findings were very similar for the three dependent variables.

The fourth section in Table 2 summarizes findings from the full model including reading level. Adding reading level to Model 3 caused the coefficient associated with race to become insignificant, suggesting that reading level explained the differences across race groups in age- and sex-adjusted episodic memory scores. These results were consistent across different composite episodic memory scores.

Sensitivity Analyses

There are a range of methods to detect and account for DIF (Millsap & Everson, 1993; Teresi, 2006) that might yield different results. We found similar results for race using single factor or bi-factor multiple indicator—multiple cause (MIMIC) models as those we report for the IRT approach. The consistency of findings across the two approaches (MIMIC *vs.* IRT) is reassuring, as is the consistency of findings when we relaxed the single factor assumption (single factor *vs.* bi-factor MIMIC).

We did not detect any multicollinearity. We assessed the variation inflation factors (VIF) for old models (age dichotomized) and new models (age centered and treated as continuous), all of the VIFs were less than 4, indicating no multicollinearity was detected. We matched participants on age, years of education, and sex to derive a sample of 546. We repeated our regression models in this matched data set and confirmed our main findings observed in Model 4 of Table 2 (see Table 3).

We performed additional regression analyses on the entire sample in which we treated age, years of education and childhood cognitive activity as continuous variables. Findings were essentially the same as our primary analyses.

We repeated analyses after excluding people whose reading test scores were close to the cutoff values. Using this approach, 67 participants were excluded from MARS and 48 from MAP. Results were very similar to those from the whole sample (Appendix 2).

We repeated the analyses of Model 4, substituting Digit Ordering for reading level. The coefficient for race in the model of the IRT score accounting for all forms of DIF was -0.16 ($p = .04$), in the model of the naive IRT scores it was -0.16 ($p = .03$), and in the model with composite *Z*-scores it was -0.10 ($p = .18$). These results suggest the ability of reading level to account for the effect of race on episodic memory is specific to reading level, because using a cognitive domain minimally correlated with reading level did not remove the effect of race.

DISCUSSION

The goal of this study was to investigate several possible explanations for lower episodic memory test scores among older African Americans compared to older Whites. Measurement bias, as identified by DIF analyses, did not explain differences across race in age- and sex-adjusted episodic memory scores. Several variables used as proxies for reserve did not explain these differences. However, we confirmed the findings of Manly and colleagues (2002, 2003, 2005) that education quality, as measured by reading level, explained differences in age- and sex-adjusted scores between African Americans and

Whites. This finding appears to be unique to reading level, as a measure of attention (Digit Ordering) did not have the same effect.

An important strength of this study is the evaluation of DIF. DIF analyses are common in educational testing, but still rare in neuropsychology. Without specific analyses, it is impossible to determine whether observed score differences across groups may be due to measurement bias or true group differences. We found that DIF was not responsible for differences in episodic memory test scores between African Americans and Whites. This finding is in contrast to DIF studies in other cognitive domains (Crane et al., 2008; Pedraza et al., 2009).

We used a hybrid IRT/OLR approach to DIF detection. There are a range of methods to detect and account for DIF (Millsap & Everson, 1993; Teresi, 2006) that might yield different results. We found similar results for race using a different DIF detection technique. The IRT approach used here relies on the assumption of unidimensionality. Methods for DIF assessment when this assumption is violated are not readily available, especially when the goal is to account for DIF with respect to a large number of covariates. We found the same item identified with DIF for race when we used single factor or bi-factor MIMIC models for episodic memory, suggesting that ignoring bi-factor structure may not be an important feature in our DIF findings.

African Americans tend to perform lower on episodic memory tests than Whites of similar age, but the differences are often due to differences in education, occupation or income (Dotson et al., 2009; Manly et al., 1998; McDougall et al., 2007; Mungas et al., 2009; Zsembik & Peek, 2001). In the current study, mean scores for some memory tests were actually higher for African Americans than Whites (Table 1), but African Americans were younger on average (Table 1). Indeed, in unadjusted analyses (Model 1 in Table 2), composite episodic memory scores did not differ across race. In adjusted analyses, African Americans had poorer age- and sex-adjusted episodic memory scores (Model 2 in Table 2). In our study, reserve factors other than reading level did not explain differences across race groups in age-adjusted episodic memory scores (Model 3 in Table 3). Reading level itself did explain differences across race groups in age adjusted episodic memory scores (Model 4 in Table 3). This effect was specific for reading level, as the race effect was still present in models that excluded reading level but included Digit Ordering.

Prior research has identified reading level as a proxy of educational quality associated with cognitive decline (Manly et al., 2002). This factor has been identified as particularly important to comparisons of neuropsychological testing results across groups of elders characterized by diverse languages and ethnic backgrounds (Cosentino et al., 2007; Manly et al., 2002). Two tests of reading level were used in the analyses: WRAT-3 and NART. MARS selected the WRAT-3 due to concerns about floor effects for the NART among minority elders. A cross-validation study found the WRAT-3 and NART to be comparable measures of premorbid intelligence (Johnstone, Callahan, Kapila, & Bouman, 1996). We are unaware of any formulas or other means of translating between the two measures. While other statistical methods (e.g., Bland-Altman plots) might prove useful to compare these tests, our sample size of 10 individuals with data from both tests was insufficient for these methods. Our categorization into high and low reading levels might be considered somewhat crude, with the distinct possibility of misclassification. The fact that this crude variable explained differences in age- and sex-adjusted episodic memory scores, while a series of other factors associated with reserve did not explain these differences, is remarkable. Our results remained unchanged when we omitted people with reading scores close to the cutoff used to distinguish between high and low scores, increasing our confidence in our findings. Results of additional sensitivity analyses in which we substituted Digit Ordering for episodic

memory further buttress the impressive nature of this finding. There was no misclassification for Digit Ordering—the same test was used in both studies—but it did not explain age- and sex-adjusted episodic memory score differences between African Americans and Whites.

As noted above, the WRAT and NART have been conceptualized as measures of reading level indicating educational quality (as we have done here) and also as measures of premorbid intelligence (Johnstone, Callahan, Kapila, & Bouman, 1996). We have used these tests in models that have already adjusted for years of education, parental education, income at midlife, and childhood cognitive activities—all factors likely also associated with intelligence but none of which explained racial differences in episodic memory scores. Furthermore, the effect of reading ability to explain racial differences was unique, as Digit Ordering did not explain racial differences in episodic memory scores. Digit Ordering is also correlated with intelligence (Luo, Chen, Zen, & Murray, 2010). While we cannot rule out the possibility that intellectual ability rather than educational quality explains differences across race in episodic memory scores, our analyses suggest that reading test scores alone—and not the other factors considered here—are able to explain these differences, suggesting that there is something unique about reading test scores not shared by these other factors.

As in any observational study, residual and/or unmeasured confounding variables may explain our findings. Unmeasured confounders (i.e., those not included in the current study) might include environmental factors (e.g., pollutants) and genetic differences. The complexities of race and culture are also unmeasured factors that may influence the performance of ethnically diverse older adults on neuropsychological tests. Aspects of culture such as acculturation contribute to older adults' performances on episodic memory tests (Manly et al., 2004).

We used somewhat crude dichotomous indicators of each factor associated with cognitive reserve in our DIF assessments and in our regression models, which raises the possibility of residual confounding. For example, based on responses to the question regarding income at age 40, we dichotomized participants into those with incomes below the median family income in the year they were 40 versus those at or above the median income. It is possible that levels of wealth well over the poverty line may not be related to additional brain protection than more modest levels of wealth, while levels of wealth close to or below the poverty line may be more linearly related to brain insults. By dichotomizing these variables, we are necessarily grouping together individuals who may nevertheless have variability in risk. When we treated the variables as continuous, our results were unchanged.

The generalizability of the results may be limited by the geographic location of the study population, the specific inclusion criteria used for the two studies, and the focus on African Americans and Whites. Furthermore, generalizability to other older African Americans may be limited by the relatively high education level in the current sample. Recall bias could possibly impact the measurement of some of our covariates such as income at age 40, childhood cognitive activity, and educational experience, though we do not expect this bias to be different across race groups. Analyses in which we matched on sex, age and years of education did not substantially change our findings. That result suggests that multiple linear regression is an adequate approach to determine the effect of race on episodic memory performance.

These results may also be limited to the specific cognitive domain, episodic memory, examined and the neuropsychological tests used to measure this domain. Indeed, Crane et al. (2008) found DIF was more important in explaining differences across race/ethnic groups for a fluency and working memory composite. The cross-sectional analyses we performed

did not allow us to comment on rates of decline of episodic memory functioning over time. Thus, we cannot comment on whether rates of decline may differ by race, or whether any such difference may be due to DIF, demographic factors, or factors associated with reserve.

In conclusion, we found on average, older African Americans had lower age- and sex-adjusted mean episodic memory scores than Whites. Those differences are not due to ignoring DIF. We tested several factors related to reserve identified from previous research, and none of these explained differences across race groups in age- and sex-adjusted episodic memory scores. However, reading level, posited to be an indicator of the quality of educational experiences, did explain differences across race groups in age- and sex-adjusted mean episodic memory scores. This finding was not generalizable to other cognitive tests. These findings reinforce prior work (Manly et al., 2002, 2003, 2005) that stressed the importance of measuring and accounting for the quality of education (as measured by reading level) in studies of older individuals from racially diverse samples.

Acknowledgments

We thank the participants in the Rush Memory and Aging Project and the Minority Aging Research Study, and the staff of the Rush Alzheimer's Disease Center. Data collection was supported by the following National Institute of Aging grants: (R01AG17917, D Bennett, PI) and (R01AG022018, L Barnes, PI). Data analyses were supported by R01AG029672 (P Crane, PI). Parts of this manuscript were presented at the National Multicultural Conference & Summit 2011 in Seattle, Washington.

References

- Albert M, Smith LA, Scherr PA, Taylor JO, Evans DA, Funkenstein HH. Use of brief cognitive tests to identify individuals in the community with clinically diagnosed Alzheimer's disease. *The International Journal of Neuroscience*. 1991; 57(3–4):167–178. Retrieved from PM:1938160. [PubMed: 1938160]
- Arvanitakis Z, Bennett DA, Wilson RS, Barnes LL. Diabetes and cognitive systems in older black and white persons. *Alzheimer Disease and Associated Disorders*. 2010; 24(1):37–42.10.1097/WAD.0b013e3181a6bed5 [PubMed: 19568148]
- Barnes LL, Wilson RS, de Leon CF, Bennett DA. The relation of lifetime cognitive activity and lifetime access to resources to late-life cognitive function in older African Americans. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*. 2006; 13(3–4):516–528. K8005180774554M6 [pii]. 10.1080/138255890969519
- Bennett DA, Schneider JA, Buchman AS, Mendes de LC, Bienias JL, Wilson RS. The Rush Memory and Aging Project: Study design and baseline characteristics of the study cohort. *Neuroepidemiology*. 2005; 25(4):163–175. NED2005025004163 [pii]. 10.1159/000087446 [PubMed: 16103727]
- Boone KB, Victor TL, Wen J, Razani J, Ponton M. The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology*. 2007; 22(3):355–365. S0887-6177(07)00017-0 [pii]. 10.1016/j.acn.2007.01.010 [PubMed: 17320344]
- Brickman AM, Cabo R, Manly JJ. Ethical issues in cross-cultural neuropsychology. *Applied Neuropsychology*. 2006; 13(2):91–100.10.1207/s15324826an1302_4 [PubMed: 17009882]
- Brunner EJ. Social and biological determinants of cognitive aging. *Neurobiology of Aging*. 2005; 26(Suppl 1):17–20. S0197-4580(05)00299-X [pii]. 10.1016/j.neurobiolaging.2005.09.024 [PubMed: 16257477]
- Camilli, G.; Shepard, LA. *Methods for identifying biased test items*. Thousand Oaks, CA: Sage; 1994.
- Cooper JA, Sagar HJ. Incidental and intentional recall in Parkinson's disease: An account based on diminished attentional resources. *Journal of Clinical and Experimental Neuropsychology*. 1993; 15(5):713–731. [PubMed: 8276931]

- Cosentino S, Manly J, Mungas D. Do reading tests measure the same construct in multiethnic and multilingual older persons? *Journal of the International Neuropsychological Society*. 2007; 13(2): 228–236. S1355617707070257 [pii]. 10.1017/S1355617707070257 [PubMed: 17286880]
- Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. *DIFdetect and difwithpar*. *Med Care*. 2006; 44(11 Suppl 3):S115–S123. 00005650-200611001-00017 [pii]. 10.1097/01.mlr.0000245183.28384.ed [PubMed: 17060818]
- Crane PK, Gibbons LE, Narasimhalu K, Lai JS, Cella D. Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. *Quality of Life Research*. 2007; 16(1):101–114.10.1007/s11136-006-0035-7 [PubMed: 17111233]
- Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, Teresi JA. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*. 2007; 16(Suppl 1):69–84.10.1007/s11136-007-9185-5 [PubMed: 17554640]
- Crane PK, Narasimhalu K, Gibbons LE, Pedraza O, Mehta KM, Tang Y, Mungas DM. Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society*. 2008; 14(5):746–759. S1355617708081162 [pii]. 10.1017/S1355617708081162 [PubMed: 18764970]
- Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*. 2004; 23(2):241–256.10.1002/sim.1713 [PubMed: 14716726]
- Dilworth-Anderson P, Hendrie HC, Manly JJ, Khachaturian AS, Fazio S. Diagnosis and assessment of Alzheimer's disease in diverse populations. *Alzheimers & Dementia*. 2008; 4(4):305–309. S1552-5260(08)00077-0 [pii]. 10.1016/j.jalz.2008.03.001
- Dorans NJ, Kulick E. Differential item functioning on the Mini-Mental State Examination. An application of the Mantel-Haenszel and standardization procedures. *Medical Care*. 2006; 44(11 Suppl 3):S107–S114. 00005650-200611001-00016 [pii]. 10.1097/01.mlr.0000245182.36914.4a [PubMed: 17060817]
- Dotson VM, Kitner-Triolo MH, Evans MK, Zonderman AB. Effects of race and socioeconomic status on the relative influence of education and literacy on cognitive functioning. *Journal of the International Neuropsychological Society*. 2009; 15(4):580–589. S1355617709090821 [pii]. 10.1017/S1355617709090821 [PubMed: 19573276]
- Everson-Rose SA, Mendes de Leon CF, Bienias JL, Wilson RS, Evans DA. Early life conditions and cognitive functioning in later life. *American Journal of Epidemiology*. 2003; 158(11):1083–1089. [PubMed: 14630604]
- Fillenbaum GG, Peterson B, Welsh-Bohmer KA, Kukull WA, Heyman A. Progression of Alzheimer's disease in black and white patients: The CERAD experience, part XVI. Consortium to Establish a Registry for Alzheimer's Disease. *Neurology*. 1998; 51(1):154–158. [PubMed: 9674795]
- Flaughner RL. The many definitions of test bias. *American Psychologist*. 1978; 33(7):671–679.
- Gasquoine PG. Race-norming of neuropsychological tests. *Neuropsychology Review*. 2009; 19(2): 250–262.10.1007/s11065-009-9090-5 [PubMed: 19294515]
- Gibbons LE, McCurry S, Rhoads K, Masaki K, White L, Borenstein AR, Crane PC. Japanese-English language equivalence of the Cognitive Abilities Screening Instrument among Japanese-Americans. *International Psychogeriatrics*. 2009; 21(1):129–137. S1041610208007862 [pii]. 10.1017/S1041610208007862 [PubMed: 18947456]
- Gurland BJ, Wilder DE, Lantigua R, Stern Y, Chen J, Killeffer EH, Mayeux R. Rates of dementia in three ethnoracial groups. *International Journal of Geriatric Psychiatry*. 1999; 14(6):481–493. [pii]. 10.1002/(SICI)1099-1166(199906)14:6<481::AID-GPS959>3.0.CO;2-5 [PubMed: 10398359]
- Johnstone B, Callahan CD, Kapila CJ, Bouman DE. The comparability of the WRAT-R reading test and NAART as estimates of premorbid intelligence in neurologically impaired patients. *Archives of Clinical Neuropsychology*. 1996; 11(6):513–519. 0887-6177(96)82330-4 [pii]. [PubMed: 14588456]

- Jones RN. Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health*. 2003; 7(2):83–102. 10.1080/1360786031000045872. 6NQP8W6NLFX55HFV [pii]. [PubMed: 12745387]
- Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination. Detecting differential item functioning using MIMIC modeling. *Medical Care*. 2006; 44(11 Suppl 3):S124–S133. 00005650-200611001-00018 [pii]. 10.1097/01.mlr.0000245250.50114.0f [PubMed: 17060819]
- Jones RN, Fong TG, Metzger E, Tulebaev S, Yang FM, Alsop DC, Inouye SK. Aging, brain disease, and reserve: Implications for delirium. *The American Journal of Geriatric Psychiatry*. 2010; 18(2): 117–127. 00019442-201002000-00004 [pii]. 10.1097/JGP. 0b013e3181b972e8 [PubMed: 20104068]
- Kaplan GA, Turrell G, Lynch JW, Everson SA, Helkala EL, Salonen JT. Childhood socioeconomic position and cognitive function in adulthood. *International Journal of Epidemiology*. 2001; 30(2): 256–263.
- Kit KA, Tuokko HA, Mateer CA. A review of the stereotype threat literature and its application in a neurological population. *Neuropsychology Review*. 2008; 18(2):132–148. 10.1007/s11065-008-9059-9 [PubMed: 18415682]
- Loewenstein DA, Arguelles T, Arguelles S, Linn-Fuentes P. Potential cultural bias in the neuropsychological assessment of the older adult. *Journal of Clinical and Experimental Neuropsychology*. 1994; 16(4):623–629. [PubMed: 7962363]
- Luo DL, Chen G, Zen F, Murray B. Modeling work memory tasks on the item level. *Intelligence*. 2010; 38(1):66–82. 10.1016/j.intell.2009.07.003 [PubMed: 20190871]
- Lord, FM.; Novick, MR. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley; 1968.
- Manly JJ. Critical issues in cultural neuropsychology: Profit from diversity. *Neuropsychology Review*. 2008; 18(3):179–183. 10.1007/s11065-008-9068-8 [PubMed: 18814033]
- Manly JJ, Byrd DA, Touradji P, Stern Y. Acculturation, reading level, and neuropsychological test performance among African American elders. *Applied Neuropsychology*. 2004; 11(1):37–46. 10.1207/s15324826an1101_5 [PubMed: 15471745]
- Manly JJ, Jacobs DM, Sano M, Bell K, Merchant CA, Small SA, Stern Y. Cognitive test performance among nondemented elderly African Americans and whites. *Neurology*. 1998; 50(5):1238–1245. [PubMed: 9595969]
- Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*. 2002; 8(3):341–348. [PubMed: 11939693]
- Manly JJ, Schupf N, Tang MX, Stern Y. Cognitive decline and literacy among ethnically diverse elders. *Journal of Geriatric Psychiatry and Neurology*. 2005; 18(4):213–217. 18/4/213 [pii]. 10.1177/0891988705281868 [PubMed: 16306242]
- Manly JJ, Touradji P, Tang MX, Stern Y. Literacy and memory decline among ethnically diverse elders. *Journal of Clinical and Experimental Neuropsychology*. 2003; 25(5):680–690. [PubMed: 12815505]
- Masel MC, Peek MK. Ethnic differences in cognitive function over time. *Annals of Epidemiology*. 2009; 19(11):778–783. S1047-2797(09)00175-6 [pii]. 10.1016/j.annepidem.2009.06.008 [PubMed: 19656690]
- McDonald, RP. *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum; 1999.
- McDougall GJ Jr, Vaughan PW, Acee TW, Becker H. Memory performance and mild cognitive impairment in Black and White community elders. *Ethnicity & Disease*. 2007; 17(2):381–388. [PubMed: 17682374]
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*. 1984; 34(7):939–944. [PubMed: 6610841]
- Millsap RE, Everson HT. Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*. 1993; 17(4):297–334.

- Morales LS, Flowers C, Gutierrez P, Kleinman M, Teresi JA. Item and scale differential functioning of the Mini-Mental State Exam assessed using the Differential Item and Test Functioning (DFIT) Framework. *Medical Care*. 2006; 44(11 Suppl 3):S143–S151. 00005650-200611001-00020 [pii]. 10.1097/01.mlr.0000245141.70946.29 [PubMed: 17060821]
- Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G. the CERAD Investigators . The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*. 1989; 39(9):1159–1165. [PubMed: 2771064]
- Mungas D, Reed BR, Farias ST, DeCarli C. Age and education effects on relationships of cognitive test scores with brain structure in demographically diverse older persons. *Psychology and Aging*. 2009; 24(1):116–128. 2009-03151-003 [pii]. 10.1037/a0013421 [PubMed: 19290743]
- Muraki, E.; Bock, RD. PARSCALE 4: IRT item analysis and test scoring for rating-scale data [computer program]. Chicago, IL: Scientific Software; 2003.
- Nelson, HE. The National Adult Reading Test (NART): Test Manual. Windsor, UK: NFER Nelson; 1982.
- Parker C, Philp I. Screening for cognitive impairment among older people in black and minority ethnic groups. *Age and Ageing*. 2004; 33(5):447–452. afh135 [pii]. 10.1093/ageing/afh135 [PubMed: 15217776]
- Pedraza O, Graff-Radford NR, Smith GE, Ivnik RJ, Willis FB, Petersen RC, Lucas JA. Differential item functioning of the Boston Naming Test in cognitively normal African American and Caucasian older adults. *Journal of the International Neuropsychological Society*. 2009; 15(5):758–768. S1355617709990361 [pii]. 10.1017/S1355617709990361 [PubMed: 19570311]
- Pedraza O, Mungas D. Measurement in cross-cultural neuropsychology. *Neuropsychology Review*. 2008; 18(3):184–193.10.1007/s11065-008-9067-9 [PubMed: 18814034]
- Ramirez M, Teresi JA, Holmes D, Gurland B, Lantigua R. Differential item functioning (DIF) and the Mini-Mental State Examination (MMSE). Overview, sample, and issues of translation. *Medical Care*. 2006; 44(11 Suppl 3):S95–S106. 00005650-200611001-00015 [pii]. 10.1097/01.mlr.0000245181.96133.db [PubMed: 17060840]
- Robertson K, Liner J, Heaton R. Neuropsychological assessment of HIV-infected populations in international settings. *Neuropsychology Review*. 2009; 19(2):232–249.10.1007/s11065-009-9096-z [PubMed: 19455425]
- Rogers MA, Plassman BL, Kabeto M, Fisher GG, McArdle JJ, Llewellyn DJ, Langa KM. Parental education and late-life dementia in the United States. *Journal of Geriatric Psychiatry and Neurology*. 2009; 22(1):71–80. 0891988708328220 [pii]. 10.1177/0891988708328220 [PubMed: 19073840]
- Rosselli M, Ardila A. The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition*. 2003; 52(3):326–333. S0278262603001702 [pii]. [PubMed: 12907177]
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*. 1969; 17
- Scarmeas N, Stern Y. Cognitive reserve and lifestyle. *Journal of Clinical and Experimental Neuropsychology*. 2003; 25(5):625–633. [PubMed: 12815500]
- Schwartz BS, Glass TA, Bolla KI, Stewart WF, Glass G, Rasmussen M, Bandeen-Roche K. Disparities in cognitive functioning by race/ethnicity in the Baltimore Memory Study. *Environmental Health Perspectives*. 2004; 112(3):314–320. [PubMed: 14998746]
- Scruggs TE, Lifson SA. Current conceptions of test-wiseness: Myths and realities. *School Psychology Review*. 1985; 14(3):339–350.
- Singh-Manoux A, Richards M, Marmot M. Socioeconomic position across the lifecourse: How does it relate to cognitive function in mid-life? *Annals of Epidemiology*. 2005; 15(8):572–578. S1047-2797(04)00323-0 [pii]. 10.1016/j.annepidem.2004.10.007 [PubMed: 16118001]
- Stern Y. Cognitive reserve and Alzheimer disease. *Alzheimer Disease and Associated Disorders*. 2006; 20(2):112–117. 00002093-200604000-00006 [pii]. 10.1097/01.wad.0000213815.20177.19 [PubMed: 16772747]

- Stern Y. Cognitive reserve. *Neuropsychologia*. 2009; 47(10):2015–2028. S0028-3932(09)00123-7 [pii]. 10.1016/j.neuropsychologia.2009.03.004 [PubMed: 19467352]
- Stern Y, Albert S, Tang MX, Tsai WY. Rate of memory decline in AD is related to education and occupation: Cognitive reserve? *Neurology*. 1999; 53(9):1942–1947. [PubMed: 10599762]
- Stern Y, Gurland B, Tatemichi TK, Tang MX, Wilder D, Mayeux R. Influence of education and occupation on the incidence of Alzheimer's disease. *The Journal of the American Medical Association*. 1994; 271(13):1004–1010.
- Stricker LJ, Emmerich W. Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement*. 1999; 36(4):347–366.
- Teng EL, Manly JJ. Neuropsychological testing: Helpful or harmful? *Alzheimer Disease and Associated Disorders*. 2005; 19(4):267–271. 00002093-200510000-00016 [pii]. [PubMed: 16327357]
- Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Medical Care*. 2006; 44(11 Suppl 3):S152–S170. 00005650-200611001-00021 [pii]. 10.1097/01.mlr.0000245142.74628.ab [PubMed: 17060822]
- Teresi JA, Golden RR, Cross P, Gurland B, Kleinman M, Wilder D. Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology*. 1995; 48(4):473–483. 0895-4356(94)00159-N [pii]. [PubMed: 7722601]
- Teresi JA, Holmes D, Ramirez M, Gurland BJ, Lantigua R. Performance of cognitive tests among different racial/ethnic and education groups: Findings of differential item functioning and possible bias. *Journal of Mental Health and Aging*. 2001; 7(1):79–90.
- Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*. 2000; 19(11–12):1651–1683. [pii]. 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AIDSIM453>3.0.CO;2-H [PubMed: 10844726]
- Teresi JA, Stewart AL, Morales LS, Stahl SM. Measurement in a multi-ethnic society. Overview to the special issue. *Medical Care*. 2006; 44(11 Suppl 3):S3–S4. 00005650-200611001-00003 [pii]. 10.1097/01.mlr.0000245437.46695.4a [PubMed: 17060831]
- Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. p. 67–113.
- Tuokko HA, Chou PH, Bowden SC, Simard M, Ska B, Crossley M. Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *Journal of the International Neuropsychological Society*. 2009; 15(3):416–425. S1355617709090602 [pii]. 10.1017/S1355617709090602 [PubMed: 19402928]
- United States Census Bureau. Income: Historical income tables - families. 2010. Retrieved from <http://www.census.gov/hhes/www/income/histinc/f07ar.html>
- Wechsler, D. *Wechsler Memory Scale - Revised Manual*. San Antonio, TX: Psychological Corporation; 1987.
- Weiner MF. Perspective on race and ethnicity in Alzheimer's disease research. *Alzheimer's & Dementia*. 2008; 4(4):233–238. S1552-5260(07)00635-8 [pii]. 10.1016/j.jalz.2007.10.016
- Whitfield KE. Challenges in cognitive assessment of African Americans in research on Alzheimer disease. *Alzheimer Disease and Associated Disorders*. 2002; 16(Suppl 2):S80–S81. Retrieved from PM:12351919. [PubMed: 12351919]
- Whitfield KE, Fillenbaum GG, Pieper C, Albert MS, Berkman LF, Blazer DG, Seeman T. The effect of race and health-related factors on naming and memory. *The MacArthur Studies of Successful Aging. Journal of Aging and Health*. 2000; 12(1):69–89. [PubMed: 10848126]
- Wilkinson, GS. *Wide Range Achievement Test 3*. Wilmington, DE: Wide Range, Inc; 1993.
- Wilson R, Barnes L, Bennett D. Assessment of lifetime participation in cognitively stimulating activities. *Journal of Clinical and Experimental Neuropsychology*. 2003; 25(5):634–642. [PubMed: 12815501]

- Wilson RS, Barnes LL, Krueger KR, Hoganson G, Bienias JL, Bennett DA. Early and late life cognitive activity and cognitive systems in old age. *Journal of the International Neuropsychological Society*. 2005; 11(4):400–407. [PubMed: 16209420]
- Wilson RS, Beckett LA, Barnes LL, Schneider JA, Bach J, Evans DA, Bennett DA. Individual differences in rates of change in cognitive abilities of older persons. *Psychology and Aging*. 2002; 17(2):179–193. [PubMed: 12061405]
- Zsembik BA, Peek MK. Race differences in cognitive functioning among older adults. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*. 2001; 56(5):S266–S274. Retrieved from PM:11522808.

APPENDIX 1: WRAT-3 AND NART ANALYSES

Note. Scatter plot of nearly simultaneous WRAT-3 and NART Z scores for participants included in both data sets. Each symbol represents one of the ten participants common to the MAP and MARS databases. Horizontal and vertical lines are provided at the cut-points used in this study: +0.48 for the WRAT-3 in MARS, and –0.98 for the NART in MAP.

APPENDIX 2: REGRESSION RESULTS EXCLUDING INDIVIDUALS WITH READING TEST Z SCORES WITHIN 0.25 OF THE CUTPOINT

	IRT score accounting for all sources of DIF			Unadjusted (“naive”) IRT score			Composite Z score		
	β	95% CI	<i>p</i> value	β	95% CI	<i>p</i> value	β	95% CI	<i>p</i> value
Model 1: Race + demographics + 5 reserve factors + Reading level									
Race	–0.06	(–0.25, 0.13)	.55	–0.05	(–0.23, 0.14)	.64	0.01	(–0.18, 0.19)	.97
Sex	–0.22	(–0.36, –0.08)	<.01	–0.27	(–0.41, –0.13)	<.01	–0.32	(–0.46, –0.18)	<.01
Age	–0.43	(–0.58, –0.28)	<.01	–0.51	(–0.66, –0.36)	<.01	–0.56	(–0.71, –0.41)	<.01
Education	0.32	(0.16, 0.48)	<.01	0.33	(0.18, 0.49)	<.01	0.34	(0.19, 0.49)	<.01

	IRT score accounting for all sources of DIF			Unadjusted ("naive") IRT score			Composite Z score		
	β	95% CI	<i>p</i> value	β	95% CI	<i>p</i> value	β	95% CI	<i>p</i> value
Father's education	-0.02	(-0.17, 0.13)	.79	-0.02	(-0.16, 0.13)	.84	0.03	(-0.11, 0.18)	.66
Mother's education	-0.09	(-0.24, 0.06)	.25	-0.10	(-0.25, 0.06)	.21	-0.13	(-0.29, 0.02)	.08
Income at age 40	0.15	(-0.04, 0.33)	.12	0.15	(-0.04, 0.33)	.12	0.12	(-0.06, 0.30)	.20
Childhood cognitive activities	0.09	(-0.05, 0.22)	.21	0.08	(-0.05, 0.22)	.22	0.08	(-0.05, 0.22)	.22
Constant	-0.39	(-0.65, -0.13)	<.01	-0.34	(-0.59, -0.09)	.01	-0.29	(-0.54, -0.04)	<.01
Reading level	0.52	(0.32, 0.72)	<.01	0.53	(0.33, 0.72)	<.01	0.52	(0.33, 0.72)	.02

Note. Regression results from a sensitivity analysis in which we omitted individuals with reading test scores close to the threshold values used to differentiate between low and high scores. The sample size reduced from 993 to 878 (67 participants were excluded from MARS and 48 participants from MAP). Regression findings are largely similar to those reported in the primary analyses.

Fig. 1.

Theoretical model of relationships between demographics, cognitive reserve and measurement bias on observed performance on episodic memory tests. Observed variables (i.e., performance on neuropsychological episodic memory tests, demographic characteristics, and indicators associated with cognitive reserve) are depicted in rectangles, while the unobserved factor (actual episodic memory functioning) is depicted in an oval. Observed episodic memory scores from neuropsychological tests (the box to the right of the figure) have only two inputs: episodic memory functioning itself (the oval at the left) and DIF. Demographics and cognitive reserve indirectly influence assessment performance through episodic memory functioning. DIF is depicted as a dashed box. DIF analyses allow us to assess whether measurement bias may be responsible for differences across groups in observed episodic memory scores. Any effect of demographics or cognitive reserve on observed performance that is *not* due to actual episodic memory functioning is depicted as a DIF effect.

Fig. 2. The derivation of study participants from the Memory and Aging Project (MAP) and Minority Aging Research Study (MARS) databases.

Table 1

Demographics, cognitive reserve, and episodic memory characteristics of sample stratified by race

Characteristic	Total sample (<i>N</i> = 993)	Black (<i>N</i> = 273)	White (<i>N</i> = 720)	<i>p</i> values
Sex				
Male: <i>n</i> (%)	291 (29)	79 (29)	212 (29)	.938
Female: <i>n</i> (%)	702 (71)	194 (71)	508 (71)	
Age				
Below or at 75 years: <i>n</i> (%)	344 (35)	184 (67)	160 (22)	<.001
Above 75 years: <i>n</i> (%)	649 (65)	89 (33)	560 (78)	<.001
Mean age (<i>SD</i>)	77.8 (7.6)	72.6 (6.01)	79.8 (7.2)	<.001
Education				
Below or at high school: <i>n</i> (%)	277 (28)	72 (26)	205 (28)	.511
Above high school: <i>n</i> (%)	716 (72)	201 (74)	515 (72)	
Mean education (<i>SD</i>)	14.8 (3.3)	15.0 (3.6)	14.8 (3.2)	.351
Father's education				
Below or at primary school: <i>n</i> (%)	477 (48)	130 (48)	347 (48)	.871
Above primary school: <i>n</i> (%)	516 (52)	143 (52)	373 (52)	
Mother's education				
Below or at primary school: <i>n</i> (%)	432 (44)	108 (40)	324 (45)	.123
Above primary school: <i>n</i> (%)	561 (56)	165 (60)	396 (55)	
Childhood cognitive activity ^a				
3 activities: <i>n</i> (%)	430 (43)	89 (33)	341 (47)	<.001
>3 activities: <i>n</i> (%)	563 (57)	184 (67)	379 (53)	<.001
Mean activity (<i>SD</i>)	3.1 (1.7)	3.1 (0.7)	3.3 (0.7)	<.001
Income at age 40 ^b				
Below or at median: <i>n</i> (%)	147 (15)	46 (17)	101 (14)	.264
Above median: <i>n</i> (%)	846 (85)	227 (83)	619 (86)	
Reading Level ^c				
Below or at cutpoint: <i>n</i> (%)	239 (24)	180 (66)	59 (8)	<.001
Above cutpoint: <i>n</i> (%)	754 (76)	93 (34)	661 (92)	
Logical Memory: mean (<i>SD</i>)				
Immediate	11.1 (4.2)	11.0 (3.9)	11.2 (4.4)	.527
Delay	9.4 (4.4)	9.2 (4.0)	9.4 (4.4)	.620
Word List Memory: mean (<i>SD</i>)				
Immediate	17.5 (4.2)	18.2 (3.8)	17.2 (4.4)	.001
Delay	5.4 (2.3)	5.7 (2.2)	5.3 (2.4)	.018
Recognition	9.6 (1.1)	9.6 (1.1)	9.5 (1.2)	.370
East Boston Naming Test: mean (<i>SD</i>)				
Memory I, immediate recall	9.5 (1.9)	9.4 (1.9)	9.5 (1.9)	.270
Memory II, delayed recall	9.1 (2.1)	9.1 (1.9)	9.1 (2.2)	.760

^aSee the Methods section for details on calculation of the childhood cognitive activity score.

^bIncome at age 40 was reported in categories of dollars. We dichotomized this variable by looking at the median family income in the U.S. for the year in which the participant was 40. See methods section for details.

^cReading level was obtained from the WRAT-3 and NART in the MARS and MAP studies, respectively. As detailed in the methods section, we analyzed data from participants in both studies to identify threshold values for the two tests that could be considered to be equivalent. The values shown in this table represent the numbers of individuals above and below those thresholds, which were a *Z* score of 0.48 for the WRAT-3 in MARS and a *Z* score of -0.94 for the NART in MAP.

Table 2

Results of regression models across Episodic Memory Scores

	IRT score accounting for all sources of DIF				Unadjusted ("naïve") IRT score				Composite Z score			
	β	95% CI	p value	β	95% CI	p value	β	95% CI	β	95% CI	p value	p value
Model 1: Race only												
Race	-0.05	(-0.19, 0.09)	.48	-0.01	(-0.15, 0.13)	.92	0.06	(-0.08, 0.20)				.37
Constant	0.01	(-0.06, 0.09)	.71	0.01	(-0.07, 0.08)	.96	-0.02	(-0.09, 0.06)				.64
Model 2: Model 1 + demographics												
Race	-0.21	(-0.37, -0.06)	.01	-0.20	(-0.35, -0.05)	.01	-0.16	(-0.31, -0.02)				.03
Sex	-0.21	(-0.35, -0.08)	<.01	-0.26	(-0.39, -0.12)	<.01	-0.31	(-0.44, -0.18)				<.01
Age	-0.36	(-0.50, -0.22)	<.01	-0.43	(-0.57, -0.29)	<.01	-0.50	(-0.64, -0.36)				<.01
Constant	0.36	(0.22, 0.49)	<.01	0.41	(0.28, 0.55)	<.01	0.46	(0.33, 0.60)				<.01
Model 3: Model 2 + all reserve factors except reading level												
Race	-0.23	(-0.38, -0.08)	<.01	-0.22	(-0.37, -0.07)	<.01	-0.18	(-0.33, -0.03)				.02
Sex	-0.25	(-0.38, -0.11)	<.01	-0.29	(-0.42, -0.16)	<.01	-0.35	(-0.47, -0.21)				<.01
Age	-0.37	(-0.51, -0.22)	<.01	-0.44	(-0.58, -0.30)	<.01	-0.50	(-0.64, -0.37)				<.01
Education	0.41	(0.26, 0.55)	<.01	0.42	(0.28, 0.57)	<.01	0.43	(0.28, 0.57)				<.01
Father's education	0.01	(-0.14, 0.15)	.95	0.01	(-0.13, 0.15)	.91	0.05	(-0.09, 0.18)				.52
Mother's education	-0.08	(-0.22, 0.07)	.29	-0.08	(-0.23, 0.06)	.26	-0.10	(-0.24, 0.04)				.17
Income at age 40	0.18	(0.01, 0.35)	.05	0.17	(-0.01, 0.34)	.05	0.13	(-0.04, 0.30)				.13
Childhood cognitive activities	0.11	(-0.02, 0.24)	.09	0.11	(-0.02, 0.23)	.09	0.10	(-0.03, 0.22)				.12
Constant	-0.09	(-0.30, 0.12)	.41	-0.04	(-0.25, 0.17)	.72	0.04	(-0.17, 0.25)				.73
Model 4: Model 3 + reading level												
Race	0.01	(-0.17, 0.19)	.92	0.02	(-0.16, 0.20)	.80	0.06	(-0.11, 0.24)				.47
Sex	-0.23	(-0.36, -0.10)	<.01	-0.27	(-0.40, -0.14)	<.01	-0.33	(-0.46, -0.20)				<.01
Age	-0.38	(-0.52, -0.24)	<.01	-0.45	(-0.59, -0.32)	<.01	-0.52	(-0.66, -0.38)				<.01
Education	0.36	(0.21, 0.50)	<.01	0.37	(0.23, 0.51)	<.01	0.37	(0.23, 0.52)				<.01
Father's education	-0.02	(-0.15, 0.12)	.80	-0.02	(-0.15, 0.12)	.83	0.02	(-0.12, 0.16)				.75
Mother's education	-0.10	(-0.24, 0.04)	.17	-0.10	(-0.25, 0.04)	.15	-0.12	(-0.26, 0.02)				.09
Income at age 40	0.16	(-0.02, 0.33)	.07	0.15	(-0.02, 0.32)	.08	0.11	(-0.06, 0.28)				.19
Childhood cognitive activities	0.09	(-0.03, 0.22)	.06	0.09	(-0.03, 0.21)	.15	0.08	(-0.04, 0.20)				.20
Constant	-0.38	(-0.63, -0.14)	<.01	-0.34	(-0.58, -0.10)	.01	-0.26	(-0.50, -0.02)				.03

\$watermark-text

\$watermark-text

\$watermark-text

	IRT score accounting for all sources of DIF			Unadjusted ("naive") IRT score			Composite Z score		
	β	95% CI	p value	β	95% CI	p value	β	95% CI	p value
Reading level	0.42	(0.24, 0.60)	<.01	0.43	(0.25, 0.61)	<.01	0.43	(0.25, 0.61)	<.01

Table 3

Regression results for the matched analyses ($n = 546$)

$N = 546$	IRT score accounting for all sources of DIF			Unadjusted ("naïve") IRT score			Composite z score		
	β	95% CI	p value	β	95% CI	p value	β	95% CI	p value
Model 1: Race only									
Race	-.24	(-.41, -.08)	<.01	-.22	(-.39, -.06)	.01	-.16	(-.33, .01)	.05
Constant	.21	(.09, .32)	<.01	.22	(.11, .34)	<.01	.21	(.09, .32)	<.01
Model 2: Model 1 + demographics									
Race	-.27	(-.43, -.11)	<.01	-.26	(-.41, -.10)	<.01	-.20	(-.36, -.04)	.01
Sex	-.19	(-.37, -.01)	.04	-.23	(-.41, -.05)	.01	-.23	(-.41, -.06)	.01
Age	-.29	(-.45, -.12)	<.01	-.36	(-.53, -.20)	<.01	-.46	(-.62, -.29)	<.01
Constant	.38	(.24, .52)	<.01	.43	(.30, .57)	<.01	.46	(.33, .60)	<.01
Model 3: Model 2 + all reserve factors except reading level									
Race	-.30	(-.46, -.14)	<.01	-.29	(-.44, -.13)	<.01	-.23	(-.38, -.07)	.01
Sex	-.19	(-.36, -.01)	.04	-.23	(-.40, -.05)	<.01	-.23	(-.40, -.06)	.01
Age	-.29	(-.45, -.12)	<.01	-.36	(-.52, -.20)	<.01	-.45	(-.61, -.29)	<.01
Education	.36	(.17, .54)	<.01	.38	(.19, .56)	<.01	.44	(.26, .62)	<.01
Father's education	.01	(-.17, .19)	.91	.01	(-.17, .19)	.88	.05	(-.13, .23)	.59
Mother's education	-.02	(-.21, .16)	.81	-.03	(-.21, .15)	.74	-.04	(-.22, .14)	.66
Income at age 40	.26	(.05, .47)	.02	.26	(.05, .46)	.02	.23	(.02, .43)	.03
Childhood cognitive activities	.15	(-.02, .32)	.08	.14	(-.02, .31)	.09	.08	(-.08, .25)	.32
Constant	-.16	(-.41, .09)	.21	-.11	(-.36, .13)	.38	-.08	(-.33, .16)	.51
Model 4: Model 3 + reading level									
Race	-.08	(-.27, .11)	.41	-.07	(-.25, .12)	.50	-.01	(-.20, .18)	.94
Sex	-.17	(-.34, .01)	.06	-.20	(-.37, -.03)	.02	-.21	(-.38, -.04)	.02
Age	-.30	(-.46, -.14)	<.01	-.37	(-.53, -.21)	<.01	-.46	(-.62, -.30)	<.01
Education	.30	(.11, .49)	<.01	.32	(.14, .50)	<.01	.38	(.20, .56)	<.01
Father's education	-.03	(-.21, .16)	.78	-.02	(-.20, .16)	.80	.01	(-.16, .19)	.88
Mother's education	-.05	(-.23, .14)	.62	-.06	(-.24, .13)	.55	-.06	(-.24, .12)	.49
Income at age 40	.24	(.03, .45)	.03	.24	(.03, .44)	.03	.21	(.01, .41)	.04
Childhood cognitive activities	.15	(-.02, .31)	.09	.14	(-.03, .30)	.10	.08	(-.08, .24)	.35
Constant	-.42	(-.70, -.14)	<.01	-.38	(-.65, -.10)	.01	-.34	(-.61, -.07)	.02

\$watermark-text

\$watermark-text

\$watermark-text

N = 546	IRT score accounting for all sources of DIF			Unadjusted (“naïve”) IRT score			Composite z score		
	β	95% CI	p value	β	95% CI	p value	β	95% CI	p value
Reading level	0.39	(0.19, 0.60)	<.01	0.40	(0.20, 0.60)	<.01	0.39	(0.19, 0.59)	<.01

Note: Overall Regression model results based on participants matched on sex, age and years of education ($n = 546$). Findings are essentially identical to regression results from the whole sample shown in Table 2.