RACE-RELATED COGNITIVE TEST BIAS IN THE ACTIVE STUDY: A MIMIC MODEL APPROACH

Adrienne T. Aiken Morgan
Michael Marsiske
Joseph M. Dzierzewski

Department of Clinical and Health Psychology, University of Florida, Gainesville, Florida, USA

Richard N. Jones

Hebrew SeniorLife, Boston, Massachusetts, USA

Keith E. Whitfield

Department of Psychology and Neuroscience, Duke University, Durham, North Carolina, USA

Kathy E. Johnson

Department of Psychology, Indiana University–Purdue University at Indianapolis, Indianapolis, Indiana, USA

Received 8 February 2007; accepted 25 April 2009.
Address correspondence to Michael Marsiske, Associate Professor, Department of Clinical and Health Psychology, University of Florida, PO Box 100165, Gainesville, FL 32610-0165, USA. E-mail: marsiske@phhp.ufl.edu
Mary K. Cresci

College of Nursing, Wayne State University, Detroit, Michigan, USA

The present study investigated evidence for race-related test bias in cognitive measures used in the baseline assessment of the ACTIVE clinical trial. Test bias against African Americans has been documented in both cognitive aging and early life span studies. Despite significant mean performance differences, Multiple Indicators Multiple Causes (MIMIC) models suggested most differences were at the construct level. There was little evidence that specific measures put either group at particular advantage or disadvantage and little evidence of cognitive test bias in this sample. Small group differences in education, cognitive status, and health suggest positive selection may have attenuated possible biases.

The goal of the present investigation was to examine whether there was evidence of race-related bias in cognitive tests used in a large trial of cognitive interventions with older adults. Although much work has been done to examine mean level differences in cognitive performance between racial and ethnic groups (e.g., Manly et al., 1998; Manly, Jacobs, Touradji, Small, & Stern, 2002; Whitfield et al., 2000), the present investigation extends the current literature by examining the extent to which such differences represent general differences at the level of the cognitive constructs of interest, or whether they also represent test-specific effects (i.e., such that some tests show specific race-group differences above and beyond general group differences at the latent level). Race in this study is operationalized as a comparison between African American and White older adults selected for participation in a cognitive clinical trial. The term African American in this study is used to characterize individuals of African, African American, and African Caribbean descent, whereas the term White is used to include persons of European descent.

Race Differences in Late-Life Cognition

Evidence from a corpus of studies on the influence of race on late-life cognition suggest mean level differences in performance on a variety of cognitive measures between African American and White older adults. Differences, usually in the form of higher performance for Whites, have been reported for tests of intelligence (e.g., Heaton, Ryan, Grant, & Matthews, 1996; Kaufman, McLean, & Reynolds, 1988; Kush et al., 2001; Vincent, 1991) and for cognitive screening
tests/batteries (e.g., Escobar, 1986; Fillenbaum, Heyman, Williams, Prosnitz, & Burchett, 1990; Inouye, Albert, Mohs, Sun, & Berkman, 1993; Whitfield et al., 2000; Zsembik & Peek, 2001; Manly et al., 1998; Manly et al., 2002; Patton et al., 2003; Unverzagt et al., 1996). This persistently lower performance by African Americans, which is present throughout the working life span (Avolio & Waldman, 1994), leads to earlier and more frequent cognitive impairment classifications and diagnoses of Alzheimer’s disease and other dementias in African American elders (e.g., Carlson, Brandt, Carson, & Kawas, 1998; Inouye et al., 1993; Manly et al., 1998; Marcopulos, McLain, & Giuliano, 1997; Ripich, Carpenter, & Ziol, 1997; Unverzagt et al., 1996, Whitfield, 2002). In light of this disparity in cognitive performance and classification between African American and White elders, several researchers have noted that the influence of certain demographic and health factors are infrequently considered in many existing published studies exploring racial group differences in cognition and should be more deliberately investigated (Whitfield et al., 2000; Izquierdo-Rorrera & Waldstein, 2002; Zsembik & Peek, 2001).

Education is commonly invoked as a major explanation for cognitive test differences between racial groups. Studies have indicated that African American adults, on average, are likely to have attained less formal education than White adults (e.g., Harper & Alexander, 1990), and these differences exist over and above cohort differences in education attainment (Adams-Price, 1993). Although most studies have considered education attainment in terms of number of years, years of education alone has been an inadequate explanation for group differences in cognitive performance; controlling for group differences in educational attainment has generally not explained the lower performance of African American elders on cognitive and neuropsychological measures (Manly et al., 1998; Manly & Jacobs, 2002). Several have argued that “years of education” does not have similar meaning across older racial groups in the United States (Jones, 2003). Furthermore, this education nonequivalence is also attributable to the historical effects of school segregation in the United States prior to 1954, and associated factors such as lower education expenditures, shorter school years, and higher student/teacher ratios that were experienced by African American students (Loewenstein, Arguelles, Arguelles, & Linn-Fuentes, 1994; Whitfield & Wiggins, 2003; Manly et al., 2002).

Health is another possible explanation for racial group differences in cognition. Health disparities in hypertension, diabetes, heart disease, and stroke are present throughout much of adulthood and
have been associated with poorer health outcomes and neuropsychological impairment (National Center for Health Statistics [NCHS], 1990; Berkman & Mullen, 1997; Whitfield, Weidner, Clark, & Anderson, 2002; Miles, 2002; Gurland, et al., 1999). One particularly important health predictor of cognition appears to be physical function, which predicts poorer performance on tasks like Boston Naming, Mini-Mental State Examination (MMSE), and Digit Symbol (Whitfield, Baker-Thomas, & Heyward, 1999, Whitfield et al., 2000; Rosano et al., 2005). Physical function may reflect both impaired performance factors (e.g., use of a pencil), as well as generally impaired health and vitality, serving as an index of underlying biological/physiological contributors to function (Marsiske, Klumb, & Baltes, 1997).

Although sociodemographic factors such as education, literacy, and health/physical function may be a partial explanation for race differences in late-life cognition, biased tests are also a possible explanation. Test bias is defined as a property of a test that causes it to be unusually sensitive to group differences. When a test is biased, it consistently leads to particularly poorer performance in one group versus another. Much of the work that has been done to understand race-related test bias has been conducted in the young, particularly children and adolescents of school age. Research has been focused on the appropriate selection of psychological test instruments in school settings (Knauss, 2003), educational and achievement testing (Banks, 2006), and language tests and corresponding normative data that may be biased against bilingual children (Saenz & Huer, 2003). Although the presence of race-related test bias has not always been found, it has been acknowledged that disproportionate representation of racial minorities in special education programs in U.S. school systems (Skiba, Knesting, & Bush, 2002) contributes to performance differences on tests.

Fewer studies have examined the influence of test bias in older populations, which motivates the present study. In one study exploring racial item bias in an older adult sample, findings suggested that there was racially related differential item functioning (DIF) for Black/African American and White participants on a modified Telephone Interview for Cognitive Status (TICS) (Jones, 2003). These differences included DIF on several TICS items (name objects, count backwards from 20, serial sevens subtraction, and name the president/vice-president). This DIF accounted for most of the mean cognitive performance group difference found, whereas background variables of low education in the Black/African American group and high income in the White group accounted for the remaining
difference (Jones, 2003). In another study examining test bias of dementia-screening instruments, findings indicated the Mini-Cog detected cognitive impairment in a racially diverse sample as well as or better than the Mini-Mental State Examination (MMSE) (Borson, Scanlan, Watanabe, Tu, & Lessig, 2005). Further, the Mini-Cog was found to be less biased by low education and low literacy (Borson et al., 2005). In a review paper examining DIF and item bias among cognitive measures used to assess the elderly, the authors concluded that many items on three dementia-screening measures (Mini-Mental Status Examination, the Short Portable Mental Status Questionnaire, and the Mattis Dementia Rating Scale) tended to yield different levels of performance across education and racial groups (Teresi, Holmes, & Ramirez, 2002). Tests that were shorter, included easier items, and relied less on language/literacy skills and more on over-learned memory skills were less sensitive to differences among individuals of varying educational and racial backgrounds (Teresi et al., 2002). This is consistent with findings by Aiken-Morgan, Marsiske, and Whitfield (2008) and Manly et al. (1998, 2002). Nevertheless, more work is needed to examine DIF in other screening and neuropsychological tests (Teresi et al., 2002).

Importantly, studies of late-life race-related bias in cognitive tests have so far been conducted primarily with neuropsychological and cognitive screening instruments. There has been a relative absence of such investigations in the normal cognitive aging literature. In part, this may reflect the low representation of persons of color in the broader cognitive aging literature. Thus, one goal of the present investigation was to expand the study of cognitive test bias to include a broader set of measures. Furthermore, the focus in this paper was not on differential item functioning, but on differential test functioning (i.e., “test bias”) to determine whether some cognitive tests show specific patterns of advantage/disadvantage for particular racial groups, after controlling for global performance groups differences. This approach is not unrelated to the idea of “differential diagnostic evidence” (Mast, MacNeill, & Lichtenberg, 2002), which suggests that some measures may be more sensitive to certain kind of group differences (e.g., vascular versus Alzheimer’s dementia) than others.

**MIMIC Models**

One way of evaluating test bias is through the use of the Multiple Indicators Multiple Causes (MIMIC) models, an application of structural equation modeling (Muthen, 1989). The present article used MIMIC modeling to investigate equivalence of ability distributions
and to measure calibrations, discriminations, and differential measure functioning/bias in African American and White older adults in the ACTIVE (Advanced Cognitive Training in Independent and Vital Elderly) trial.

The MIMIC approach investigates whether a given covariate or set of covariates has a unique and direct effect on a given measure of a construct. A schematic of this approach is shown in Figure 1. Figure 1 displays a construct, an exogenous covariate, and three measures of the construct. In the MIMIC model, factor loadings

![Figure 1. MIMIC Model: An illustration.](image)

*Note.* a paths are factor loadings, and express the relationship between the latent construct and each of three measures of that construct; the b path shows the relationship between the exogenous covariate and the latent construct. This path addresses whether there is an overall relationship between the covariate and the construct. The c path assesses whether, after controlling for the relationship between covariate and the construct, there is an *additional* unique direct residual relationship between the covariate and each indicator. If the relationship between the covariate and Measure 1 is *more strongly positive* than the relationship between the covariate and the construct, then one would expect an additional significant direct positive path between the covariate and Measure 1. This would suggest that Measure 1 is positively biased. In other words, it shows a significantly stronger relationship to the covariate than the overall construct. Expressed differently, this would mean that the residual variance in Measure 1 (variance not explained by the common construct factor) shows additional relationship to the covariate, above and beyond that covariate’s indirect influence on Measure 1 via the latent construct.
(relationship between the measures and their construct) are estimated. In addition, the covariate is allowed to have direct effects on both the latent construct, and the specific measures/indicators of that construct. If a given measure has no unique relationship with the covariate, then all of its relationship with the covariate will be fully mediated by the latent construct. On the other hand, if the relationship between the covariate and a measure is stronger than this purely indirect path would imply, then one would require an additional path between the covariate and the measure to capture this additional relationship. The requirement of this additional path would suggest bias; that is, the measure shows a significantly stronger relationship to the covariate than expected based on its saturation by the latent construct. Expressed differently, this would mean that the residual variance in the measure (variance not explained by the common construct factor) shows additional relationship to the covariate, above and beyond that covariate’s indirect influence on the measure via the latent construct. To concretize these ideas, if the covariate were “race,” the presence of an additional significant path from “race” to a measure would suggest that measure is more strongly related to “race” (i.e., evinces stronger race differences) than the construct to which it belongs. The same logic applies in the opposite direction; a significant path would also result when a given measure is less related to the covariate (e.g., race) than the indirect relationship via the construct would imply. The MIMIC model will be considered again, more explicitly, when the specific model to be evaluated in this study is discussed in the Methods section of this paper.

It should be noted that test bias can be discussed in different ways. One way, the approach used in the current study, explores threats to the validity of a test that can occur in the scoring of individuals on underlying traits (this paper). A second approach to studying bias refers to the differential prediction of outcomes (i.e., in the use and interpretation of a test’s trait estimate), but that is not the focus of the current paper. The current study focused on bias as a concept internal to the trait.

A few prior studies have used MIMIC to examine differential item functioning and test bias in research and clinical samples. As mentioned above, Jones (2003) used MIMIC modeling to demonstrate DIF for a Black/African American sample on a modified TICS measure. Additionally, Mast and Lichtenberg (2000) used MIMIC to examine the effect of population heterogeneity, but not race, among geriatric patients on factor structure and DIF on a measure of functional independence (functional independence measure; FIM). Their findings suggested three motor functioning and three cognitive
functioning items showed DIF systematically across sample subgroups (young-old versus old-old, male versus female, and depressed versus nondepressed), and the authors concluded that scores on the FIM may lead to biased comparisons of functional abilities across subgroups (Mast & Lichtenberg, 2000).

Finally, Mast et al. (2002) demonstrated the MIMIC methodology by integrating neuropsychological test data to a MIMIC model to examine the influence of cerebrovascular disease (CVD) on global cognitive impairment dimensions and individual tests, after controlling for cognitive impairment. Their findings demonstrated the presence of CVD in dementia was unrelated either to dimensions of global cognitive impairment or to performance on 9 of 10 neuropsychological tests (Mast et al., 2002). Although Mast and colleagues did not consider race, they demonstrated the practical utility of applying the MIMIC approach to investigate possible sources of bias in commonly used gerontological and geriatric measures. The current study examines race-related cognitive differences, at both the latent and test-specific levels, using MIMIC (Muthen, 1989) modeling.

**Specific Aims**

The present investigation extends the current literature by addressing two specific aims. First, we explored differences in mean cognitive test performance between African American and White community-dwelling older adults. Second, we used MIMIC modeling to specifically investigate racial group differences and possible test bias.

**METHODS**

**The ACTIVE Study**

The ACTIVE study is a randomized, controlled, single-masked clinical trial designed to examine whether cognitive training can affect cognitively based measures of daily functioning (Ball et al., 2002; Jobe et al., 2001). Data for the current study are taken from the ACTIVE baseline assessment, which was completed prior to the beginning of the cognitive intervention part of the trial. A variety of recruitment strategies (e.g., through local churches and senior organizations) were employed by each of six field sites (Birmingham, Alabama; Boston, Massachusetts; Indianapolis, Indiana; Baltimore, Maryland; Detroit, Michigan; and north central Pennsylvania), and are documented in Jobe et al. (2001).
Participants

There were a total of 2802 participants in the sample; 61 of these older adults were excluded from the present analyses because either they did not identify themselves as African American/Black or European American/White or were incorrectly randomized, and thus represented categories of insufficient frequency to permit group comparisons. The present analyses were based on the remaining subsample of 2741 adults aged 65 and older, of whom 75.8% were women, with a mean age of 73.6 years and a mean of 13.5 years of education (Table 1). The sample was comprised of approximately 26.6% \((n = 729)\) African Americans. The African American group was significantly younger \((p < .001)\) than the White group (mean age = 72.1 years versus 74.1 years, respectively) and included significantly more females \((p < .001; 83.2\% \text{ versus } 73.2\%, \text{ respectively})\). African Americans also reported having significantly fewer years of education \((p < .001; \text{ mean education} = 13.0 \text{ versus } 13.7 \text{ years, respectively})\) and had poorer self-rated physical function on the SF-36 (Ware & Sherbourne, 1992) \((p < .001; \text{ mean} = 65.0 \text{ versus } 70.2)\). For the SF-36 scale, we compared groups on all subscales and found the physical function subscale to be the only physical health subscale to differ significantly by race. Additionally, the SF-36 physical function subscale was the most related to cognition in this sample.

<table>
<thead>
<tr>
<th>Table 1. Characteristics of the ACTIVE sample at baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Total sample ((N = 2741))</td>
</tr>
<tr>
<td>African Americans ((n = 729))</td>
</tr>
<tr>
<td>Whites ((n = 2012))</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Years of education</td>
</tr>
<tr>
<td>Gender (% female)</td>
</tr>
<tr>
<td>Physical function</td>
</tr>
</tbody>
</table>

\(^a\)Degrees of freedom and \(t\)-statistic were adjusted for nonhomogeneity of variance.

Note. \(t\) tests show significance of the comparison between African American and White participants; for Gender, a corresponding chi-squared test was used. No covariate adjustment was employed.
Exclusion Criteria
Potential participants in the ACTIVE study were screened to exclude individuals based on criteria related to seven factors: (1) age <65 years at initial screening; (2) probable dementia, as defined by a score of 22 or lower on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), a widely used screening tool, and/or a self-reported diagnosis of Alzheimer’s disease; (3) substantial functional decline, as determined by self-reported need for assistance in performing activities of daily living (ADLs) related to bathing, dressing, or personal hygiene; (4) specific medical conditions, such as stroke, in the past 12 months, certain high-fatality cancers (e.g., liver, lung, esophagus), or current chemotherapy and or radiation treatment for any cancer; (5) severe sensory loss; (6) communication difficulties; and (7) recent or current participation in cognitive training studies; or unavailability during any phase of the study. The sample was not restricted to people born in the United States, nor was it restricted to persons for whom English was the first language. Trained telephone screeners were asked to score whether participants could understand and make themselves understood, as part of the screening process. The rationale for these exclusion criteria related to the goals of overall clinical trial study is reviewed by Jobe et al. (2001).

Measures and Data Collection Procedures
Demographic data were collected via telephone interview. Cognitive data were collected over the course of two 90-min testing sessions conducted individually and a third 120-min session conducted in a small group. Age, gender, years of education, and physical function were used as covariates in all analyses to control for nonequivalent sampling by race (i.e., African Americans were younger.

1The criterion of MMSE score ≥23 was used to ensure adequate recruitment of African American older adults to meet the goals of the pilot and overall ACTIVE clinical trial. It should be noted that MMSE cutoff was selected because it was lower than some published recommended cutoff scores for the MMSE (Crum, Anthony, Bassett, & Folstein, 1993), and was thought to be more broadly inclusive. In addition, the “one-size-fits-all” cutoff (Cullen et al., 2005) was developed to permit quick implementation during the initial screening phase of the study (i.e., testers quickly used it to determine initial eligibility in the field, so a more nuanced set of cutoffs divided by age, education and/or race was not feasible). The consequence of this cutoff did lead to significant race differences in exclusion due to the MMSE. Of the 316 Whites who were excluded, 115 (36.4%) received an MMSE below 23; of the 222 blacks who were excluded, 150 (67.6%) were excluded on the basis of an MMSE below 23. This represented a significant difference (with continuity correction, χ²(N=538, df=1) = 49.46, p <.001).
Table 2. Cognitive measures

<table>
<thead>
<tr>
<th>Domain</th>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reasoning</td>
<td>Word Series Test; Schaie-Thurstone Adult Mental Abilities Test (Schaie, 1985)</td>
<td>Assessed the ability to correctly determine how a series of words (e.g., days of the week, months of the year) progress. There were 30 multiple-choice items.</td>
</tr>
<tr>
<td></td>
<td>Letter Series Test; Schaie-Thurstone Adult Mental Abilities Test (Schiae, 1985)</td>
<td>Assessed the ability to correctly identify the letter that appropriately completes a series of letters for each of the test items. There were 30 multiple-choice items.</td>
</tr>
<tr>
<td></td>
<td>Letter Sets Test; Kit of Factor Referenced Cognitive Tests (Ekstrom, French, Harman, &amp; Derman, 1976)</td>
<td>Assessed the ability to identify the rule that related four groups of letters to each other, and subsequently identify a fifth group of letters to which the rule did not apply. There were 15 items.</td>
</tr>
<tr>
<td>Memory</td>
<td>Rey Auditory Verbal Learning Test (Rey, 1964)</td>
<td>Assessed the ability to form and retain new verbal memories for a list of 15 semantically unrelated words. There were five trials of learning and recall.</td>
</tr>
<tr>
<td></td>
<td>Hopkins Verbal Learning Test—Revised; (Benedict Schretlen, Groninger, &amp; Brandt, 1998)</td>
<td>Assessed the ability to form and retain new verbal memories for a list of 12 semantically related words. There were three trials of learning and recall. A recognition trial was also given, but is not included in this study.</td>
</tr>
<tr>
<td></td>
<td>Rivermead Behavioural Memory Test (Wilson, Cockburn, &amp; Baddeley, 1985)</td>
<td>Assessed the ability to learn and retain verbal memories for a brief story of 21 propositions. A free recall trial was given.</td>
</tr>
<tr>
<td>Processing speed</td>
<td>Useful Field of View Test (UFOV) (Ball &amp; Owsley, 1993)</td>
<td>Assessed elements of an individual’s field of visual attention: simple perceptual identification (UFOV1: Speed), perceptual identification and localization of peripheral target (UFOV2: Divided Attention), identification and localization of peripheral object in visual clutter (UFOV3: Selective Attention), and perceptual comparison and peripheral localization in clutter (UFOV4: Selective Attention). Scores on each subtask represented the minimum presentation time needed for 75% accuracy in central and peripheral tasks.</td>
</tr>
</tbody>
</table>
more likely to be female, and had poorer physical function) (Table 1). Insufficient data were available at the ACTIVE baseline assessment to permit us to include other covariates, such as income and socioeconomic status.

The cognitive battery consisted of measures of reasoning, speed, and memory. The chosen measures represent commonly used cognitive instruments to test cognitive performance in aging populations. In addition, the measures were selected to reflect the endpoints of the larger ACTIVE clinical trial. Baseline assessments of memory, reasoning, and speed of processing, which were the planned targets of training for the full study, were included in the present study, as delineated in Table 2. Variables are coded such that higher scores mean better performance for reasoning and memory measures, whereas lower scores on speed measures indicate better performance.

RESULTS

Analysis Plan

All structural equation models in this study were estimated using AMOS 16.0 program (Arbuckle, 2007). Structural equation models employed Blom-transformed cognitive scores, producing more normally distributed measures (Ball et al., 2002; Blom, 1958). Models were evaluated using overall fit indexes representing the root mean square error of approximation (RMSEA), a fit index indicating the discrepancy between the original and reproduced covariance matrix divided by the degrees of freedom and a fit index for which values of .08 or lower are indicative of adequate fit. Comparative fit index (CFI), incremental fit index (IFI), and the normed fit index (NFI) were also examined (Bentler, 1989; Bentler & Bonet, 1980; Bollen, 1989; Marsh, Balla, & McDonald, 1988), and were thought to be indicative of adequate fit when they assumed values of 0.95 or higher. Chi-square statistics were not examined as indicators of specific model fits, because of known problems of inflation in larger samples and with deviations from normality in variables under study (Akaike, 1987; Carmines & McIver, 1981). At the same time, to compare nested models, a chi-square difference test was conducted. For nested models, the difference between the model chi-square statistics is distributed as a chi-square statistic, with the difference in degrees of freedom ($df$) as the $df$ of the difference statistic. In addition, Akaike’s Information Criterion (AIC) was examined; better models should evince smaller AIC values.
Mean Cognitive Performance Differences

To explore differences in mean cognitive test performance between community-dwelling African American and White older adults, group means were compared, using t test analyses. These comparisons were done to determine whether significant differences in performance existed between race groups on the selected cognitive measures, controlling for demographic variables and data collection site. As indicated in Table 3, there were significant mean differences between the African American and White groups, reflecting better performance for Whites, for all cognitive measures. For several measures, there was nonhomogeneity of variance, thus t values and degrees of freedom for these t test comparisons were adjusted for unequal variances.

Ability Factor Model and Invariance

Before testing the MIMIC model, including race group differences at both the level of the cognitive factors and specific tests, the general fit of the proposed three-factor model of cognition (reasoning, memory, speed of processing) was examined. A confirmatory factor analysis, using AMOS 16.0, was conducted. The planned MIMIC model required a combined model across both race groups, so that a dummy variable representing race could be included as a predictor in the model. However, to evaluate the feasibility of combining the data from both racial groups, we first estimated our cognitive factor model while examining factor invariance across the two groups (models were estimated using covariance metric). Thus, the three-factor cognitive ability structure was estimated simultaneously in the African American and White samples, and the fit in both groups was evaluated. Invariance tests examined the effects on fit of three different levels of invariance constraint (see Table 4). For all models tested, uniqueness terms (variance in each measure not explained by its factor) were allowed to vary between groups in all models. A partially invariant solution, with loading invariance between groups, represented the best model and also had the lowest levels of AIC.

Completely standardized factor loadings for the combined groups have been included in Table 5. In the White sample, correlations between the cognitive factors were as follows: reasoning and memory, $r = .66$; reasoning and speed, $r = -.66$; memory and speed, $r = -.58$. In the African American sample, correlations between the cognitive factors were as follows: reasoning and memory, $r = .63$; reasoning and speed, $r = -.53$; memory and speed, $r = -.47$. 
Table 3. Mean cognitive test performance of the ACTIVE sample at baseline (controlling for demographic variables and data collection site)

<table>
<thead>
<tr>
<th></th>
<th>Total sample (N = 2741)</th>
<th>African Americans (n = 729)</th>
<th>Whites (n = 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Word Series</td>
<td>9.49</td>
<td>4.89</td>
<td>7.21</td>
</tr>
<tr>
<td>Letter Series</td>
<td>10.04</td>
<td>5.55</td>
<td>7.17</td>
</tr>
<tr>
<td>Letter Sets</td>
<td>5.75</td>
<td>2.79</td>
<td>4.87</td>
</tr>
<tr>
<td>AVLT</td>
<td>48.51</td>
<td>10.59</td>
<td>45.92</td>
</tr>
<tr>
<td>HVLT</td>
<td>26.11</td>
<td>5.49</td>
<td>24.64</td>
</tr>
<tr>
<td>Rivermead</td>
<td>6.29</td>
<td>2.75</td>
<td>5.39</td>
</tr>
<tr>
<td>UFOV1</td>
<td>30.81</td>
<td>40.76</td>
<td>34.01</td>
</tr>
<tr>
<td>UFOV2</td>
<td>133.28</td>
<td>124.86</td>
<td>156.91</td>
</tr>
<tr>
<td>UFOV3</td>
<td>320.64</td>
<td>134.16</td>
<td>341.54</td>
</tr>
<tr>
<td>UFOV4</td>
<td>456.75</td>
<td>68.55</td>
<td>464.80</td>
</tr>
</tbody>
</table>

Note. AVLT = Rey Auditory Verbal Learning Test; HVLT = Hopkins Verbal Learning Test; Rivermead = Rivermead Paragraph Recall Task; UFOV1–UFOV4 = Useful Field of View subtasks 1 through 4.

*Degrees of freedom and t-statistic were adjusted for nonhomogeneity of variance.
Next, MIMIC modeling examined whether there was specific race-related test bias in the ACTIVE baseline cognitive battery. MIMIC models allowed for the influence of multiple factors (e.g., race, age, gender, education, physical function) to be considered when determining whether measures of cognitive factors function similarly

**Table 4. Invariance tests for proposed three-factor model of cognition**

\( N = 2741; n = 729, \) African Americans; \( n = 2012, \) Whites

<table>
<thead>
<tr>
<th>Model</th>
<th>( \chi^2(df) )</th>
<th>( \chi^2 ) Difference</th>
<th>( p )</th>
<th>NFI</th>
<th>IFI</th>
<th>RMSEA</th>
<th>CFI</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model One</td>
<td>444.17 (77)</td>
<td>—</td>
<td>&lt;.001</td>
<td>0.96</td>
<td>0.97</td>
<td>0.042</td>
<td>0.97</td>
<td>550.17</td>
</tr>
<tr>
<td>Model Two</td>
<td>421.34 (71)</td>
<td>22.83 (6)</td>
<td>0.0009</td>
<td>0.97</td>
<td>0.97</td>
<td>0.042</td>
<td>0.97</td>
<td>539.34</td>
</tr>
<tr>
<td>Model Three</td>
<td>410.67 (64)</td>
<td>10.67 (7)</td>
<td>0.154</td>
<td>0.97</td>
<td>0.97</td>
<td>0.044</td>
<td>0.97</td>
<td>542.67</td>
</tr>
</tbody>
</table>

*Note.* In Model One, both unstandardized factor loadings and factor covariances were constrained to equality across groups. In Model Two, only unstandardized factor loadings were constrained to equality across groups. In Model Three, configural invariance (i.e., same pattern of factor loadings in both groups) was tested.

NFI = Normed Fit Index; IFI = Incremental Fit Index; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index.

**MIMIC Model Results**

Next, MIMIC modeling examined whether there was specific race-related test bias in the ACTIVE baseline cognitive battery. MIMIC models allowed for the influence of multiple factors (e.g., race, age, gender, education, physical function) to be considered when determining whether measures of cognitive factors function similarly

**Table 5. Completely standardized factor loadings of cognitive measures on hypothesized latent factors at baseline assessment occasion**

\( N = 2741; n = 729, \) African Americans; \( n = 2012, \) Whites

<table>
<thead>
<tr>
<th>Reasoning</th>
<th>Memory</th>
<th>Processing speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Series</td>
<td>.91***</td>
<td></td>
</tr>
<tr>
<td>Letter Series</td>
<td>.91***</td>
<td></td>
</tr>
<tr>
<td>Letter Sets</td>
<td>.69***</td>
<td></td>
</tr>
<tr>
<td>AVLT</td>
<td>.79***</td>
<td></td>
</tr>
<tr>
<td>HVLT</td>
<td>.83***</td>
<td></td>
</tr>
<tr>
<td>Rivermead</td>
<td>.59***</td>
<td></td>
</tr>
<tr>
<td>UFOV1</td>
<td>.53***</td>
<td></td>
</tr>
<tr>
<td>UFOV2</td>
<td>.73***</td>
<td></td>
</tr>
<tr>
<td>UFOV3</td>
<td>.83***</td>
<td></td>
</tr>
<tr>
<td>UFOV4</td>
<td>.61***</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* AVLT = Rey Auditory Verbal Learning Test; HVLT = Hopkins Verbal Learning Test; Rivermead = Rivermead Paragraph Recall Task; UFOV1–UFOV4 = Useful Field of View subtasks 1 through 4. **p < .001.

\( ***p < .001. \)
An extension of confirmatory factor analysis, a MIMIC model is a structural equation model (SEM) that has one or more latent variables simultaneously identified by multiple endogenous indicators and multiple exogenous causal variables (Muthen, 1989). Endogenous indicators in this paper are the tests that compose a latent cognitive factor, whereas exogenous indicators are demographic variables of race, age, gender, education, and physical function.

A schematic of the estimated three-factor MIMIC model is shown in Figure 2. In this figure, paths are shown only for one latent cognitive factor for simplicity. Paths from each latent cognitive variable (reasoning, memory, and speed) to its corresponding cognitive tests (path $e$) reflect factor loadings. MIMIC models were run in two steps. In the first step, paths were estimated to the latent construct from race (path $a$), while simultaneously controlling for exogenous demographic covariates (age, gender, education, physical function, as well as testing site [not shown]; path $b$). In a second model, paths to the

![Figure 2. MIMIC Model in the current study.](image)

Note. MIMIC model of race effects on reasoning, memory, and speed, with years of education, age and gender as covariates. Paths shown are as follows: $a$, influence of ethnicity on latent variable; $b$, influence of covariates on latent variable; $c$, influence of race on specific measure; $d$, influence of covariates on specific measure; $e$, factor loading (relationship between latent construct and specific measure). **Paths are shown only for one latent factor, but these paths were estimated for all latent variables and measured indicators.** Dummy variables representing each of the assessment sites were also included as exogenous covariates in these models, but they are not shown in order to simplify the figure.
latent constructs were fixed (to the values obtained in the preceding model), and paths were estimated from race to each specific test (path c), controlling for the effects of the demographic covariates on that test (path d). The models were run in these two steps to permit identification, which is akin to procedures used in extension analysis (Kemper & Summer, 2001; Schaie et al., 2005). Note that this model employs a combined model, collapsed across race groups. This combined model was justified based on (1) earlier findings of two-group factor invariance in loadings, (2) the fact that a highly constrained model with both factor loadings and covariances forced to equality also fit well, and (3) the technical requirement of MIMIC models to have a combined solution; most MIMIC model studies do not first test invariance (Muthen, 1989). It should be noted that for dichotomous variables, such as race, the path coefficient expresses the mean difference between groups; that is, what is the unit difference in the dependent variable (i.e., specific measure, construct) associated with the difference between groups.

The logic of the MIMIC test is as follows. Race differences on a given test (e.g., Letter Series) may be due to two sources. One source is because the underlying construct that produces performance on the Letter Series test (i.e., Reasoning) itself evinces race differences. Thus, race differences at the construct level reflect themselves in race differences on the specific measures that represent that construct. However, if there is additional variance in the Letter Series test that is unique to that measure (i.e., not reflective of the underlying common factor; in other words, the “unique” or “residual” variance in Letter Series), and if that unique variance in Letter Series is also related to race, then there may be an additional, direct relationship between race and the Letter Series. What this would mean is that the Letter Series test evinces race differences for two reasons: (a) reasoning shows race differences, and (b) the specific test contains further race differences above and beyond the reasoning construct. This additional race-difference variance in the Letter Series is the bias. It says that Letter Series shows more race differences than we would expect on the basis on the Reasoning construct alone.

Each cognitive construct was modeled in two steps. The first step, which estimated regression coefficients between race and the cognitive factors, controlling for covariates (race, age, gender, education, and physical function) and data collection site is summarized in Table 6; coefficients shown are standardized regression weights. All regression coefficients (ranging between .05 and .50 in magnitude) were statistically significant, except for the coefficients between
gender and the speed \((b = .00)\) factor. Of special interest were the regression coefficients for the effect of race on the latent cognitive factors, which were statistically significant and of low to moderate magnitude. Standardized \(b\)-weights were reasoning \(= -.36\), memory \(= -.26\), and speed \(= .15\). The fit of these models were acceptable: Reasoning \(\chi^2(20) = 64.46, p < .001; NFI = 0.99; IFI = 0.99; RMSEA = 0.028; CFI = 0.99;\) Model AIC = 232.46; Memory \(\chi^2(20) = 309.81, p < .001; NFI = 0.96;\) IFI = 0.96; RMSEA = 0.073; CFI = 0.96; Model AIC = 477.81; Speed \(\chi^2(32) = 376.76, p < .001; NFI = 0.95; IFI = 0.96; RMSEA = 0.063; CFI = 0.96;\) Model AIC = 550.76.

Next, in the second step, the existence of specific residual paths between race and the individual tests, after controlling for covariates, site, and the effects of the predictors on the latent construct, was examined. Positive coefficients suggested race group differences at the test level were less than expected based on group differences at the factor level. Table 7 displays the standardized regression coefficients obtained from these models. There were small but significant residual performance differences for African American participants on two measures: the Rivermead Behavioral Memory scale (Paragraph Recall) \((-0.05)\), and UFOV Subtest 2 \((0.05)\) \((p < .05;\) Table 7). For both tests, these paths indicated tiny test-specific performance disadvantages for African Americans. That is, the magnitude of the residual race difference was more than that of the overall latent level difference attributable to race. The fit of these models was acceptable: Reasoning \(\chi^2(2) = 0.90, p > .05; NFI = 1.00; IFI = 1.00;\) RMSEA = 0.000; CFI = 1.00; Model AIC = 204.90; Memory \(\chi^2(2) = 1.72, p > .05; NFI = 1.00;\) IFI = 1.00; RMSEA = 0.000; CFI = 1.00; Model AIC = 205.72; Speed \(\chi^2(5) = 220.40, p < .001; NFI = 0.97; IFI = 0.97;\) RMSEA = 0.125; CFI = 0.97; Model AIC = 448.40. Because these second steps were not nested
within the ones above (the second steps fixed elements of the preceding solutions, so the models could be identified), hierarchical model tests could not be conducted. Nonetheless, these results indicated that above the expected race group performance difference at the level of the latent constructs of reasoning, memory, and speed, there were are also test-specific differences that showed a slight consistent pattern of disadvantage for African American elders in this sample. These significant effects were nonetheless very weak (Cohen, 1988), and their significance may in part be attributable to the substantial statistical power afforded by the relatively large size of the multisite sample.

**DISCUSSION**

The goal of the present study was to examine the presence of race-related test bias in the ACTIVE study sample of African American and White older adults. Test bias in the present context was defined as a property of a test that caused it to be unusually sensitive to group differences and would consistently lead to poorer performance in one racial group versus another. Although various sources of racial group differences have been examined in previous research (e.g., education, literacy, physical function), little work to date has examined the role of test bias in understanding group
differences. Results of this study found typical racial group mean differences on various measures of reasoning, speed, and memory. Further, racial group differences were also found at the latent construct level. Regarding cognitive factor invariance, analyses supported strict invariance of factor loadings across groups.

Using MIMIC modeling, we found little evidence of test-specific bias; for most measures, residual effect of race on specific tests was near zero, and only reached significance in two cases. In both cases (one involving memory and one for speed), the residual effect of race was less than 0.10 and likely reached significance only due to the substantial statistical power of this study. However, the pattern of race effects was consistent; the residual effect of African American status was positive.

Thus, there are several conclusions that can be made based on these findings. First, as has been documented consistently, racial differences in cognitive performance are clearly present in late life. Next, unlike early in the life span, or even other studies of late-life mental status, there was little evidence of test-specific bias here. This suggests that over and above important potential mediating factors (i.e., education, literacy, physical function, and socioeconomic status) that have been found to explain cognitive performance differences between African Americans and Whites in previous work, individual cognitive tests added little to no added bias for African Americans, which is an encouraging and positive finding.

Nevertheless, this study has several limitations that are important to consider. The ACTIVE study was not designed explicitly for the purpose of cognitive comparison between African American and White older adults. In addition, using a selective convenience sample means that generalizability to the broader population of older adults requires further consideration. Although essentially a sample of convenience, our sample of African Americans and Whites were reasonably well matched on mean level of education. In the United States, it is almost certain that racial discrimination made access to education highly discrepant across African American and White groups (Whitfield & Wiggins, 2003). The African Americans in this study, who on average completed 13 years of education, are less typical overall of African American elders. According to the 2000 U.S. Census, for the cohort of adults aged 65 and older in 2000, the median education for African Americans was between 9 and 11 years. In contrast, the Whites in the present sample averaged 13.7 years of school, which is more typical of the median education of White Americans in this cohort (U.S. Census Bureau, 2002). We therefore acknowledge the possibility that the selection process and the level
of mean education for this study lessened the likelihood of finding cognitive test performance bias in this sample.

We expect our findings would have yielded greater evidence of test bias had we included more individuals with lower levels of education attainment, specifically in the African American group. Further, it is a concern that bias might have been larger in a sample that better represented individuals of lower education. This is a potential limitation and requires follow-up study in a more educationally heterogeneous sample to see whether the cognitive test bias varies by educational level. Nonetheless, the present study could be taken as a "natural experiment" with a sample of positively selected African Americans in terms of educational attainment.

There remains the need to understand and explain why persistent mean and latent level racial group differences remain. These differences have implications for late-life impairment rates as well as implications for the long-term effects of early intervention attempts to reduce educational and health disparities. This assertion may extend to the cognitive status screening, given that only participants with MMSE scores of 23 or higher were admitted into the study. Some investigators have argued that different MMSE cutoffs should be used for African American and White elders (Bohnstedt, Fox, & Kohatsu, 1994). The decision of the ACTIVE investigators not to implement differential cutoffs, which was done both to address field logistics (participants needed to be screened in "on the spot") and to ensure a common minimum performance baseline in all participants who might later be randomized to a training condition, potentially might have further led to disproportionately positive selection bias for the African American group. On the other hand, from the perspective of quasiexperimental methodology, the sample selection resulted in two race groups with fewer gross differences than in the population, permitting a more careful evaluation of the unique effect of race, if any, on test bias.

Speculatively, the MMSE cutoff of 23 essentially trimmed the lower portion of the cognitive status distribution. Moreover, given the aforementioned known race differences in the MMSE distribution, it is likely that this cutoff disproportionately removed African Americans. Indeed, a total of 3357 ACTIVE African American and White participants were screened with the MMSE. Of these, 95% of White participants (2279) and 84% of African American participants (811) were MMSE eligible. This difference in eligibility was significant ($\chi^2(1) = 107.78, p < .0001, \theta = 0.18$). Beyond the obvious generalizability limitations that this finding implies, there is no way
(absent a new data collection) to determine whether bias findings would have been larger if the lower-MMSE groups could have been included. At the same time, there is little extant empirical reason to believe that bias, if it exists, only exists in the “lower portion” of the cognitive ability distribution. For now, however, the present findings seem to permit the assertion, with some statistical power, that there is little evidence of race-related bias in the top 85% (i.e., cognitively unimpaired segment) of the aging population. Nonetheless, the constrained generalizability of the present sample is again acknowledged as a limitation, even as the sample is more heterogeneous than most in the cognitive aging literature.

Adequate data on socioeconomic status (SES) were not available for the ACTIVE baseline assessment to permit analysis of the possible influences of this variable; thus, SES could not be controlled for in this study. The present study therefore cannot determine the extent to which our findings were influenced by uncontrolled SES effects. Additionally, the data were collected to measure fluid abilities (e.g., information processing), rather than crystallized abilities that are more closely related to education and acculturation and found to be more biased in assessment. Measures of fluid ability are thought to be more related to biological changes that occur with age, which might be reflected in health differentials. Thus, it is not known whether inclusion of measures of crystallized knowledge in this sample would have resulted in greater evidence for test bias.

In sum, these results suggest few reasons not to use common cognitive measures with African American community-dwelling elders. Nevertheless, additional research is essential for determining whether test bias presents more of a problem when older adults are in poor health, have very little education, or are socioeconomically vulnerable. It is therefore necessary to examine a broader selection of cognitive instruments and in a sample that is more diverse and more representative of the lower end of the health, education, and SES spectrums. Group differences then must arise from other covariates not included in this analysis. Individual factors such as literacy and health status probably account for the lower mean scores found for African Americans (Manly et al., 1998, 2002; Whitfield, 1996; Whitfield & Willis, 1998). The inclusion of other covariates is needed to fully comprehend the origin and source of group differences in cognitive aging. Finally, although it was not the focus of the current study, results of the MIMIC models demonstrated significant test bias against women on the Letter Sets test. Thus, future inquiry should also explore other potential sources of cognitive test bias in older adults.
REFERENCES


