Journal of the International Neuropsychological Society (2016), **22**, 66–75.

doi:10.1017/S1355617715001113

Examination of the Factor Structure of a Global Cognitive Function Battery across Race and Time

Lisa L. Barnes,[1,2,3] Futoshi Yumoto,[4,5] Ana Capuano,[1,2] Robert S. Wilson,[1,2,3] David A. Bennett,[1,2] AND Rochelle E. Tractenberg[4,6]

[1]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois
[2]Department of Neurological Sciences, Rush University Medical Center, Chicago, Illinois
[3]Department of Behavioral Sciences, Rush University Medical Center, Chicago, Illinois
[4]Collaborative for Research on Outcomes and Metrics, USA
[5]Merkle, Columbia, Maryland
[6]Departments of Neurology and Biostatistics, Bioinformatics & Biomathematics, Georgetown University Medical Center, Washington, DC

(RECEIVED March 3, 2015; FINAL REVISION September 24, 2015; ACCEPTED October 13, 2015; FIRST PUBLISHED ONLINE November 13, 2015)

Abstract

Older African Americans tend to perform more poorly on cognitive function tests than older Whites. One possible explanation for their poorer performance is that the tests used to assess cognition may not reflect the same construct in African Americans and Whites. Therefore, we tested measurement invariance, by race and over time, of a structured 18-test cognitive battery used in three epidemiologic cohort studies of diverse older adults. Multi-group confirmatory factor analyses were carried out with full-information maximum likelihood estimation in all models to capture as much information as was present in the observed data. Four different aspects of the data were fit to each model: comparative fit index (CFI), standardized root mean square residuals (SRMR), root mean square error of approximation (RMSEA), and model χ^2. We found that the most constrained model fit the data well (CFI = 0.950; SRMR = 0.051; RMSEA = 0.057 (90% confidence interval: 0.056, 0.059); the model $\chi^2 = 4600.68$ on 862 df), supporting the characterization of this model of cognitive test scores as invariant over time and racial group. These results support the conclusion that the cognitive test battery used in the three studies is invariant across race and time and can be used to assess cognition among African Americans and Whites in longitudinal studies. Furthermore, the lower performance of African Americans on these tests is not due to bias in the tests themselves but rather likely reflect differences in social and environmental experiences over the life course. (*JINS*, 2016, *22*, 66–75)

Keywords: Measurement invariance, African American, Longitudinal, Epidemiology, Cognition, Cohort study

INTRODUCTION

It is well documented that older African Americans tend to perform more poorly on cognitive function tests than older Whites, even after adjustment for years of education (Manly, 2005). Literacy, a commonly used marker for quality of education typically assessed by performance on reading tests, has often been found to attenuate, but not completely eliminate, the differences (Manly, Jacobs, Touradji, Small, & Stern, 2002), suggesting that other factors may play a role in the poorer performance among African Americans (e.g., Barnes, Lewis, et al., 2012). Some investigators have argued that neuropsychological tests may be biased against racial/ethnic minority populations because they were developed with and standardized on the majority white population (Arnold, Montgomery, Castaneda, & Longoria, 1994; Loewenstein, Arguelles, Arguelles, & Linn-Fuentes, 1994). In fact, there is a growing body of evidence that suggests that background variables such as strategies, cognitive styles, and familiarity with testing may vary by race/ethnicity as well (e.g., Early et al., 2013; Jones, 2003), potentially influencing interpretation of performance.

Another potential explanation for the poorer performance of older African Americans on neuropsychological tests is that the underlying latent structure of the tests themselves differs as a function of race. That is, the tests used to assess cognition may not reflect the same underlying construct in different population groups. Although equivalence of latent cognitive structure (i.e., *measurement invariance*) is critical for comparative research with neuropsychological testing,

Correspondence and reprint requests to: Lisa L Barnes, Rush Alzheimer's Disease Center, Rush University Medical Center, Armour Academic Center, 600 S. Paulina, Suite 1038, Chicago, IL 60612. E-mail: lbarnes1@rush.edu

relatively few studies test for equivalence in comparative studies, whether over age, or time, race, or other groupings. Measurement invariance is observed when, "…under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (p. 117; Horn & McArdle, 1992). To interpret differences in mean scores across groups, the mean values over the groups must represent the *same attribute* (Bollen & Curran, 2006). In this case, "different conditions" could refer to independent groups or the same group over time. The validity and utility of representing a complex measurement process within a single theoretically and empirically supported model are critically dependent on the demonstration of *invariance* of that model over time and across groups.

Measurement invariance is a characteristic of any measurement tool, and has its modern conceptualization arising originally from concerns regarding performance on admissions tests by different populations (Meredith & Teresi, 2006). Since the mid-1980s, measurement invariance has been examined more widely, particularly in the cognitive aging literature with older populations. For example, some studies have examined the measurement invariance of cognitive abilities across age (e.g., Bowden, Weiss, Holdnack, & Lloyd, 2006; Hertzog & Schaie, 1986; Schaie, Willis, Jay, & Chipuer, 1989), gender (Maitland, Intrieri, Schaie, & Willis, 2000), neurological impairment (Hayden et al., 2011; Siedlecki, Honig, & Stern, 2008), and language (Siedlecki et al., 2010; Tuokko et al., 2009). Although several studies have examined racial differences in cognitive function and cognitive decline among older adults, including change within specific domains of function (e.g., Brewster et al., 2014; Early et al., 2013; Schwartz et al., 2004; Sloan & Wang, 2005), relatively few have examined measurement invariance specifically as a function of race in late-life (e.g., Blankson & McArdle, 2013; Jones, 2003; Mungas, Widaman, Reed, & Tomaszewski, 2011). This is essential, as the model and test scores must accurately reflect the particular latent traits measured by the test *across* different populations and/or time to make valid interpretations of differences or apparent changes in performance (Horn & McArdle, 1992).

Invariance is not an all-or-none characteristic; it falls on a continuum. A "hierarchy" of invariance moves from *strict* invariance, which involves the same model form, factor loadings, residual variances for items and factors, and intercepts for all models under consideration; to *strong* invariance, relating to equivalence in model form and intercepts (not loadings); to *weak* (metric or pattern) invariance, which requires the same model form and factor loadings for all models (Meredith & Teresi, 2006). Invariance in *model form* means that the same (multiple) observed variables represent the same latent construct at each timepoint (Bollen & Curran, 2006). Invariant *factor loadings* means that, given the same model form and scaling, the observed variables relate to (load on) the latent variable(s) the same way over time. Invariance in *intercepts* of the observed items or variables/indicators means that the items or indicators are functioning in the same way to represent the latent variable invariantly over time. These invariance characteristics pertain to the fit features of the model either to the same group over time or to independent samples at a single time point. Models that can be replicated across independent samples *and* over time improve science (Mulaik, 2010), making invariance an early, important, step in modeling and measurement. Invariance of the residual variances, which is a defining characteristic of "strict" invariance, is often difficult to obtain (Chen, 2007). Invariance up to the level of intercepts ("strong-plus") permits interpretable comparisons of *latent mean differences* across groups, whereas invariance up to the level of residual variances means that group differences on the indicators of the factors are comparable (Chen, Sousa, & West, 2005). A focus on the latent means ("strong-plus"), rather than on the individual observed test scores ("strict"), means that potential sources of bias in the observed test scores (e.g., age, education), which are the indicators or items of the latent factors in our model, are de-emphasized.

The purpose of the current analysis, therefore, was to examine the degree to which the latent factor structure of a well-established cognitive battery (Wilson et al., 2002) is invariant across race (African American *vs*. White) and over time. Data come from three longitudinal community-based cohort studies:the Minority Aging Research Study (MARS), the Rush Memory and Aging Project (MAP), and the Religious Orders Study (ROS). The studies have similar recruitment strategies and operational components, and use the same cognitive battery, which facilitates merging the data. This project tested the hypothesis that the 18 cognitive tests that are common to all three longitudinal studies, which have been reported elsewhere to define five cognitive domains (Wilson et al., 2002) within a single structural equation (latent measurement) model of "cognition," are invariant over race and over time.

METHOD

Participants

Data for this project were obtained from participants enrolled in one of three community-based cohort studies of aging and cognition:MARS, MAP, or ROS. The Institutional Review Board of Rush University Medical Center approved all studies and all participants signed written informed consent. Given essentially identical recruitment techniques and a large overlap of identical data collection (including cognitive testing) across all three studies, data were merged to examine measurement invariance across race. In the current analyses, all self-reported African Americans (from all three cohorts) were grouped together, and compared to Whites (from MAP and ROS).

MARS enrolls older African Americans without known dementia living in the community, who agree to annual clinical evaluations (Barnes, Shah, Aggarwal, Bennett, & Schneider, 2012). From the start of enrollment in August 2004, 477 persons enrolled in the study, of whom 466 completed a baseline clinical evaluation. Of these, 13 met criteria for dementia at baseline (see Clinical Evaluation, below) and

https://doi.org/10.1017/S1355617715001113 Published online by Cambridge University Press

were excluded from the current study. MAP enrolls older persons of all races without known dementia from retirement communities and senior subsidized housing facilities across the Chicago metropolitan area (Bennett, Schneider, Buchman, et al., 2012). Participants agree to annual clinical evaluations and donation of their brain, spinal cord, and selected nerves and muscles at time of death. From the start of enrollment in 1998, 1459 persons enrolled and 1450 completed a baseline clinical evaluation. Of these, 79 met criteria for dementia and were excluded from the current study. ROS enrolls older Catholic nuns, priests, and brothers, from more than 40 groups across the United States (Bennett, Schneider, Arvanitakis, & Wilson, 2012). Participants are without known dementia and agree to annual evaluation and brain donation. From the start of enrollment in January 1994, 1108 persons enrolled in the study, of whom 1106 completed a baseline clinical evaluation. Of these, 76 met criteria for dementia at baseline and were excluded from the current study.

Materials and Procedure

Clinical evaluation

Participants underwent uniform, structured, clinical evaluations annually, which included a detailed medical history, neurologic examination, and neuropsychological testing, as previously described (Barnes, Shah, et al., 2012; Bennett, Schneider, Arvanitakis, et al., 2012; Bennett, Schneider, Buchman, et al., 2012). On the basis of this evaluation, an experienced clinician classified subjects with respect to Alzheimer's disease and other neurologic conditions following criteria set forth by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (McKhann et al., 1984).

Other variables

Each participant was asked to self-report their racial category, based on questions used by the U.S. Census Bureau (1990).

Cognitive tests

A battery of 19 cognitive function tests was administered in a 1-hr session. The Mini-Mental State Examination (MMSE) was used to describe the overall cognitive functioning of the cohorts, but not in analyses. The remaining 18 performance-based tests assess the levels of episodic memory, semantic memory, working memory, perceptual speed, and visuospatial abilities. Details of the cognitive function tests have been reported previously (Wilson et al., 2002, 2005). There were seven tests of episodic memory: immediate and delayed story recall of story A from the Logical Memory subtest of the WMS-R (Wechsler, 1987) and of the East Boston Story (Albert et al., 1991), and Word List Memory, Word List Recall, and Word List Recognition from the procedures established by CERAD (Morris et al., 1989); two tests of semantic memory: a 15-item version of the Boston Naming Test (Wilson et al., 2005; Kaplan, Goodglass, & Weintraub, 1983) and semantic Verbal Fluency from CERAD (Wilson et al., 2005; Morris et al., 1989); three tests of working memory: Digit Span Forward and Digit Span Backward from the Weschler Memory Test-R (Wechsler, 1987) and Digit Ordering (Cooper & Sagar, 1993); four measures of perceptual speed: Symbol Digit Modalities Test (Smith, 1982), Number Comparison (Ekstrom, French, Harman, & Kermen, 1976), and two indices from a modified version of the Stroop Neuropsychological Screening Test: the number of color names correctly read aloud in 30 s minus the number of errors, and the number of colors correctly named in 30 s minus the number of errors (Trenerry, Crosson, DeBoe, & Leber, 1989); and two tests of visuospatial ability: a 15-item version of Judgment of Line Orientation (Benton, Sivan, Hamsher, Varney, & Spreen, 1994) and a 16-item version of Standard Progressive Matrices (Raven, Court, & Raven, 1992). We created a global composite with all tests, and composite measures of episodic memory, semantic memory, working memory, perceptual speed, and visuospatial ability, by converting raw scores on each component to z scores, using the baseline mean and *SD* in the entire cohort, and then averaging the *Z* scores, as previously described (Wilson et al., 2002, 2005).

Data analysis

Multi-group confirmatory factor analyses (CFA) were carried out using MPlus 6.12 (Muthén & Muthén, 1998) with full-information maximum likelihood (FIML) estimation in all models (Allison, 2003) to capture as much information as is present in the observed data. In all cases, we sought to test the hypothesis that the model was invariant over time and race *via* CFA, and so we fit a fully constrained model first, relaxing constraints iteratively from fully constrained ("strong-plus," which includes factor structure, indicator loadings, and indicator intercepts each constrained to be equal), to measurement constrained ("strong," i.e., releasing constraints on the intercepts, but not loadings or factor structure), to structural ("weak," i.e., same items loading on same factors without the values of the loadings or intercepts constrained), to fully *un*constrained (all factors, loadings, and parameters in the set of models estimated separately). We did not constrain factor or indicator residual variances to be equal in any model. We did not plan to release any specific constraints but to move from level to level of invariance by releasing all constraints consistent with the next level when a more-constrained version failed to fit. Residuals and modification indices were never used to identify constraints to release, because our focus was on testing the hypothesis that the model was invariant, and determining the extent to which this invariance could be said to exist for this model. Instead, the analysis plan was to examine residuals to identify how or where the hypothesis that the model *is* invariant was not supported. We did not plan to change features of the model based on residuals or modification indices, but to characterize failures of invariance according to these residuals in case any were informative about lack of fit.

We used a factor-analytically derived model that was described in a previous publication (Wilson et al., 2002), but

https://doi.org/10.1017/S1355617715001113 Published online by Cambridge University Press

which has never been tested for its invariance over time or race. Model fitting specifics are provided below. We first fit this model to the data all together at baseline to ensure that the tests of measurement invariance over race and time would proceed using a well-fitting model. Fit statistics (described below) were used to study this model as well; residuals and modification indices were also used at this step only to better understand any sources of misfit in this overall measurement model. For the invariance modeling, we fit the most-constrained version of this structural equation model for cognitive measurement first (releasing constraints as described above) *within* racial groups (White and African American), separately over time. We proceeded to the point where a well-fitting model (representing the supported level of invariance) over time was identified for each racial group. Once we had iterated the CFA within racial group over time, we then started again with samples together (i.e., testing invariance across racial groups) at baseline, at year 1, and at year 2. Finally, we planned to follow the same iterative process to fit a single CFA for invariance over both time and across racial group, moving from the most- to the least-constrained version.

Four different aspects of fit to the data were assessed for *each* model: model χ^2, a general data-model fit statistic; comparative fit index (CFI), representing incremental model fit relative to independence—the closer to 1.0 the better, with acceptable models having CFI > .96; standardized root mean square residuals (SRMR) summarizing mean absolute value of the covariance residuals—the smaller (and <0.09) the better; and root mean square error of approximation (RMSEA) representing the error in the approximation of the data by the model—acceptable models have an upper bound on the 90% RMSEA confidence interval (CI) <0.06 (Hu & Bentler, 1999). Hu and Bentler (1999) propose that an acceptable model has *all* of the following fit characteristics: CFI > 0.96; and SRMR < 0.09; and RMSEA < 0.06. When we were comparing more- to less-constrained models, in addition to observing that all fit indices were in the required ranges, we also required that the CFI values change less than 0.01 (Byrne, 2006, 2011) between more- and less-constrained models. This way, we could be confident of both the fit of the model to the data and also that the incremental model fit was not changing in practically important but possibly not statistically significant ways.

Once we had our base model, the invariance testing began with the most constrained model, and as soon as a model fit the data well according to all fit indices and (if applicable) the CFI values having changed by less than 0.01, the modeling procedure stopped and the level of constraints in *this* model was taken as the level of invariance. In the event that a more- and a less-constrained model seemed to fit comparably, we planned to use a likelihood ratio test, using the difference in numbers of parameter estimates as the degrees of freedom, to determine if the less-constrained model was a significantly better fit than the more constrained model. If it was not, then the more-constrained model was retained. In every case, we planned to release all constraints consistent with the next level of invariance, that is, we did not plan to release constraints on any specific path but rather, planned to move from strong-plus to strong; from strong to weak; and from weak to no invariance—stopping as soon as a good fitting model was identified.

In addition to examining and comparing fit indices outlined above, we planned to examine the residuals (comparing the observed and model-implied variance-covariance matrices) and modification indices for two competing levels of invariance to identify, where relevant, those indicators for which the models did not fit well over time, across cohorts, or both (Tractenberg, Aisen, Weiner, Cummings, & Hancock, 2006).

RESULTS

Table 1 lists demographic characteristics across the three study cohorts; other than racial composition, characteristics were similar across cohort. Within the combined sample, African Americans were slightly younger [73.0 (*SD* = 6.8) *vs*. 78.4 (*SD* = 7.4] years) and had fewer years of education [14.9 (*SD* = 3.5) *vs*. 16.3 (*SD* = 3.5)] compared with Whites. However, baseline MMSE score 27.8 (*SD* = 2.4) for African Americans *versus* 28.2 (*SD* = 1.9) for Whites and distribution of women (76.6% in African Americans *vs*. 71.4% in Whites) were similar across the two groups.

All data at the baseline visit were used to estimate the "base" model to be tested for invariance within groups over time and across groups at each visit. Table 2 shows the sample sizes that were used in the modeling. Because the sample size for the African Americans dropped off significantly after the 2nd year of follow-up, our analyses focused on the three annual evaluations in the baseline–2nd year time frame.

The five-factor model that has been reported previously (Wilson et al., 2002) was fit to the data to ensure that our conceptual model would be an appropriate starting point. This pre-specified model fit the baseline data from all participants together well [CFI = 0.961; SRMR = 0.037; RMSEA = 0.052 (90% CI [0.049, 0.055]) with model χ^2 = 1129.467 on 122 df], once covariances were added (based on modification indices) among the errors associated with 3 of the 18 cognitive tests (Story A from the Logical Memory subtest; East Boston Story recall; and Word List Recall). This was the model (see Figure 1) we used for the invariance analyses.

Table 1. Demographic characteristics by study cohort

	MARS *n* = 453	MAP *n* = 1,371	ROS *n* = 1,030
Mean age (*SD*)	73.6 (6.3)	80.1 (7.2)	75.1 (7.2)
Mean education (*SD*)	14.8 (3.4)	14.7 (2.9)	18.3 (3.2)
% female	74	74.5	69.3
% African American	100	6.9	7.1
Mean MMSE[a] (*SD*)	27.8 (2.4)	27.9 (2.1)	28.5 (1.6)

[a]Mini-Mental State Examination.
MARS, Minority Aging Research Study; MAP, Rush Memory and Aging Project; ROS, Religious Orders Study.

https://doi.org/10.1017/S1355617715001113 Published online by Cambridge University Press

Table 2. Sample sizes, across racial groups, over time

	Baseline	Year 1	Year 2
White	2392	2120	1898
African American[a]	652	479	426
Total	1520	1303	1151

[a]African American participants in Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) were combined with those in the MARS cohort; ROS and MAP Whites were combined for the other group.

Invariance over Time

Table 3 shows the fits of the most-constrained model within each group separately to test invariance of the model over time. Because this most-constrained model fit within required ranges of all fit indices, we stopped at this level and did not fit less-constrained versions.

Our methods specified that we would stop releasing constraints as soon as we hit the model that met fit index criteria and where the CFI value did not change from the previous model (where possible) by 0.01 or more. The model fit information shown in Table 3 shows that the most-constrained model, where factor loadings, indicator intercepts *and* factor structure were all fixed to be equal within one race group at each of three annual assessments, fit the data well in both groups. As was specified earlier, covariances among the latent factors and between errors on any test score; the residual variances on the 18 tests; latent variable means; and latent variable variances were not constrained to be equal within a group over time. The most constrained model generally fit well over time within each group. The fit is marginally better for the larger sample (Whites) than for the smaller, African American sample, and examination of the modification indices and the residual matrices suggested no specific areas of misfit to address in the

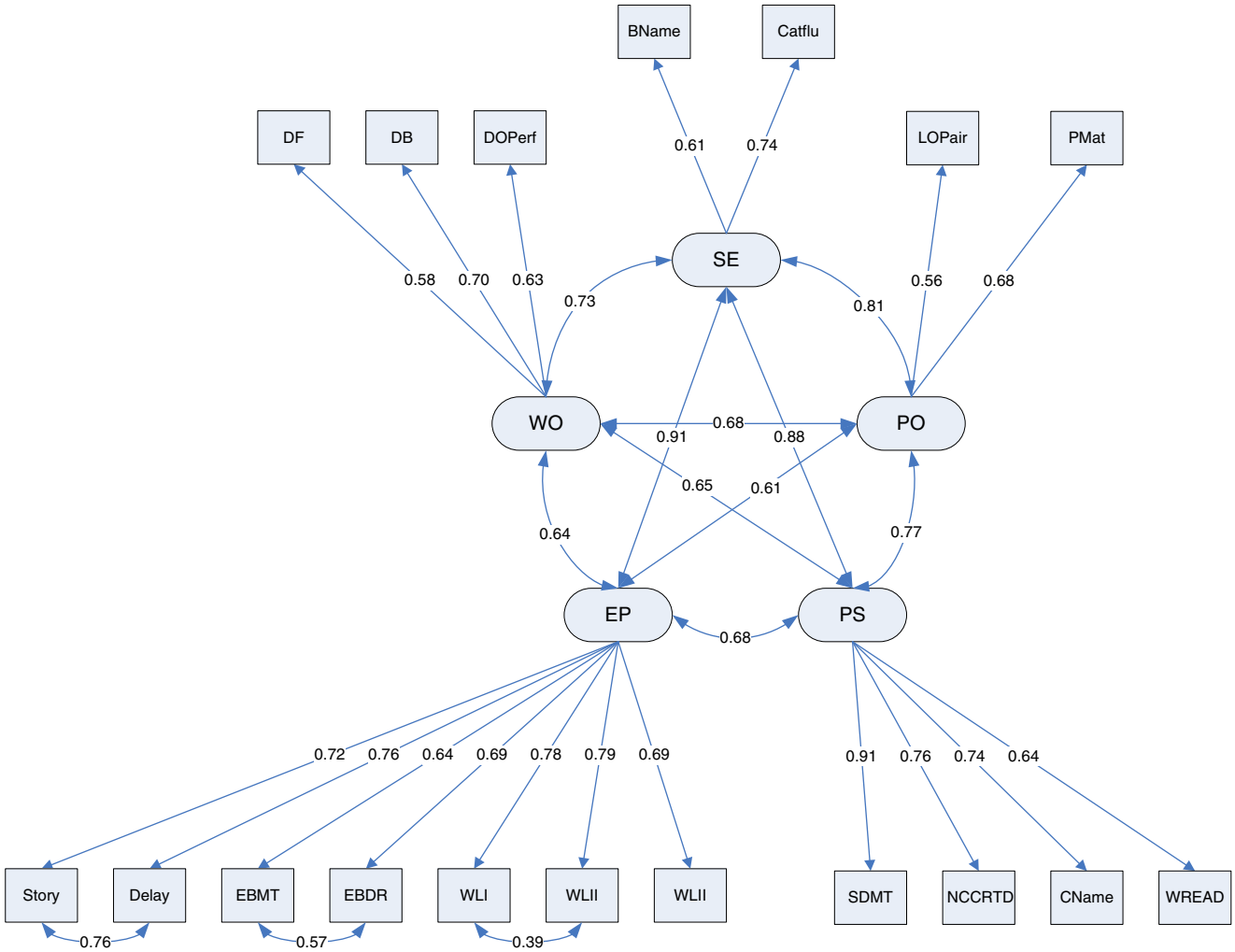

Fig. 1. Model used for the invariance analyses. Abbreviations: DF = Digits forward; DB = Digits backward; DOPerf = Digits Ordering; Bname = Boston Naming Test; Catflu = Category Fluency; LoPair = Line orientation; Pmat = Progressive Matrices; Story = Story A Immediate Recall; Delay = Story A Delayed Recall; EBMT = East Boston Memory Test; EBDR = East Boston Delayed Recall; WLI = Word List Learning; WLII = Word List Recall; WLIII = Word List Recognition; SDMT = Symbol Digits Modality Test; NCCRTD = Number Comparisons Test; Cname = Stroop, color naming test; WREAD = Stroop, word reading test. SE = Semantic Memory; PO = Visuospatial Abilities; PS = Perceptual Speed; EP = Episodic Memory; WO = Working Memory.

https://doi.org/10.1017/S1355617715001113 Published online by Cambridge University Press

Table 3. Fit of most-constrained invariance in the five factor model *over time* (baseline, 1 year, 2 year) within racial group

Fit index: how did most constrained model fit?	White	African American
Model χ^2 (418 df)	2944.295	1022.370
CFI	0.959	0.952
SRMR	0.039	0.049
RMSEA	.053 (.051, .055)	.053 (.049, .057)

CFI: comparative fit index; SRMR: standardized root mean square residuals. RMSEA: root mean square error of approximation.

model for either group. The most constrained model that was fit was a good fit by all our fit indices within each group, so less-constrained versions of this model were not attempted. These results support the characterization of *invariance of the measurement model* over time within each group.

Invariance across Groups

Table 4 shows the fits of the most constrained model (equivalent factor loadings, indicator intercepts *and* factor structure for each racial group) to the data from African American and White participants at each visit (baseline, year 1, year 2).

The model fit results in Table 4 show that, when analyzed with our five-factor model constrained to be equivalent across racial group, the most constrained model was a good fit to the data at each of the three visits. The upper bound on the 90% CI around the RMSEA value exceeded the recommended limit [0.06] (Hu & Bentler, 1999) in years 1 and 2, so we examined the modification indices and residuals matrices. There was no evidence of specific contributions to misfit in the residuals matrices. Modification indices suggested that the most theoretically and clinically plausible sources of this specific aspect of misfit involved three of the test scores (indicators) for the episodic memory (EP) factor. Given the critical contribution of episodic memory assessment in the definition and detection of cognitive aging, we re-fit this most-constrained model with all of the EP indicators' loadings unconstrained. That is, to determine if the equivalence constraint on loadings for the indicators on the EP factor was driving the too-wide range in the RMSEA 90% CI, we fit the same model, but without the constraints of equivalence on the EP factors. The fit was marginally statistically significantly better by likelihood ratio test ($\chi^2_{30} = 44.48$; $p = .043$), but Akaike's Information Criterion (AIC) favored the most-constrained model over the one without equivalent loadings on the EP factor indicators. Since AIC is not sensitive to sample size (and this likelihood ratio test is well known to be so), we conclude that overall, these results generally support the characterization of the measurement model as *invariant* across these two racial groups at each of the three annual assessments.

Table 4. Fit of most-constrained invariance in the five factor model *for racial group:* participants in the two racial groups from all three cohorts were all modeled together, separately by year

Fit index: how did most constrained model fit?	Baseline	Year 1	Year 2
Model χ^2 (270 df)	1490.4	1452.36	1408.74
CFI	0.953	0.953	0.952
SRMR	0.051	0.050	0.046
RMSEA	.054 (.052, .057)	.058 (.055, .061)	.060 (.057, .063)

CFI: comparative fit index; SRMR: standardized root mean square residuals. RMSEA: root mean square error of approximation.

Given the results in Tables 3 and 4 that the model with the same (highest) level of constraints fit within group over time (Table 3), and across groups at each time point (Table 4), we proceeded to test the fit of the model to the two groups over time. This multigroup CFA tested the theoretical - general - invariance ("strong-plus": equivalent factor loadings, indicator intercepts *and* factor structure) of this five factor model of cognitive functioning based on these 18 tests over time and racial group.

The most constrained model, with factor structure, indicator loadings, and indicator intercepts each constrained to be equal across racial groups *and* over time, was a good fit to the data (CFI = 0.950; SRMR = 0.051; RMSEA = 0.057 (90% CI [0.056, 0.059]; the model = 4600.68 on 862 df). All fit indices were within recommended range, suggesting that the measurement model, shown in Figure 1, is *invariant* across these racial groups and over time.

The final model is shown in Figure 1; the final standardized parameter estimates are included in the Figure. Table 5 presents the latent factor means.

The observation of good fit of this model to the data enables us to estimate "true" group means and explore their differences across cohorts over time. Table 5 shows the standardized mean factor scores from our final model (shown in Figure 1), estimated for racial group and time. These factor score means are standardized, so that the overall average is zero, and standard deviations are equal to one. The values in Table 5 show that, while the estimated group means for African Americans are lower in nearly every case than they are for Whites, all group values are *well* within one standard deviation of the mean (zero). That is, even if these particular groups differ in their means at any timepoint, relying on an invariant model from which factor scores can be reliably derived (as in Table 5) permits us to discern that the groups are not "significantly different" in their mean levels of any cognitive factor at any of the three annual exams. With a factor score that has been shown to be invariant to time and racial group, the observed differences—whether or not they are statistically significant—can be interpreted with confidence. Now that we have demonstrated the time and group invariance of the measurement models, these models may be used in future studies in the analysis of other outcomes like neuropathology, MRI volumetrics, or the evaluation of intra-individual variability.

https://doi.org/10.1017/S1355617715001113 Published online by Cambridge University Press

Table 5. Standardized mean factor scores from invariant model over time, by racial group

Racial group	Year	Episodic memory	Semantic memory	Working memory	Perceptual speed	Visuospatial ability
White	BL	0[a]	0[a]	0[a]	0[a]	0[a]
African American	BL	–0.043	–0.246	–0.477	–0.275	–0.870
White	1	0.068	–0.006	0.069	–0.006	–0.003
African American	1	0.156	–0.158	–0.441	–0.261	–0.825
White	2	0.059	–0.078	–0.009	–0.093	–0.050
African American	2	0.118	–0.198	–0.412	–0.193	–0.749

[a]MEAN for standardized scores = 0; SD for standardized scores is 1. White baseline scores were arbitrarily selected to be the reference group and their mean at baseline was constrained to be zero.

DISCUSSION

The present study examined the measurement invariance of a performance-based neuropsychological test battery, used in three ongoing longitudinal epidemiologic cohort studies, across race and time. Starting with the published measurement model for cognitive functioning (Wilson et al., 2002), we iteratively tested the hypothesis that this model was invariant over time within racial group, across racial group by visit, and finally, for both racial group and time within a single multi-group confirmatory analysis. In each case, we found that the most-constrained ("strong-plus") model fit the data well, supporting the characterization of this model of cognitive test scores as invariant over time and racial group. These results support the conclusion that the cognitive test battery used in MARS, MAP, and ROS is invariant across race and time and can be used to assess cognition among African Americans and Whites in our longitudinal studies. As noted, invariance up to the level of intercepts ("strong-plus") permits interpretable comparisons of *latent mean differences* across groups (Chen et al., 2005).

We focused our analyses on invariance at the level of latent factor means ("strong plus), rather than on the individual observed test scores ("strict"), for two reasons. First, cognition is a complex construct, requiring assessment of many complementary domains (i.e., the full 18-test battery). Thus, invariance on the individual test scores is not of interest in our longitudinal work, whereas the latent factors represent cognition at a sufficiently comprehensive level. Second, emphasis on the factor, rather than the individual, scores attenuates the likelihood of bias in the observed test scores (e.g., possibly arising from age, education, or other factors that we could and did not test) in group comparisons. This study of measurement invariance tested the hypothesis that the model was invariant and was, therefore, more confirmatory in nature than many other examples in the literature. We took the modeling approach articulated by Burnham and Anderson (2002), treating the effects of time and race as potential sources of bias (as described by Millsap, 2011).

In all cases, we fit a more-fully constrained model first, planning to relax constraints iteratively from this level of constraint ("strong-plus": factor structure, indicator loadings, and indicator intercepts each constrained to be equal), to measurement constrained ("strong," i.e., releasing constraints on the intercepts, but not loadings or factor structure), to structural ("weak," i.e., same items loading on same factors without the values of the loadings or intercepts constrained), to fully unconstrained ("none," all factors, loadings, and parameters in the set of models estimated separately). Because this type of modeling specifically sought to test the hypothesis of measurement invariance, releasing constraints stopped as soon as a model fit the data well. In the cases where a less-constrained model was also attempted, likelihood ratio testing of these nested models showed the less-constrained model did not fit the data statistically significantly better than the more-constrained version did, so the more constrained models were retained in all cases.

Previous studies have examined measurement invariance of cognitive abilities across age (Bowden et al., 2006; Hertzog & Schaie, 1986; Schaie et al., 1989), gender (Maitland et al., 2000), neurological impairment (Hayden et al., 2011; Siedlecki et al., 2008), and language (Siedlecki et al., 2010; Tuokko et al., 2009). Relatively few studies, however, have examined measurement invariance as a function of race (Blankson & McArdle, 2013; Jones, 2003; Mungas et al., 2011). Measurement invariance as a function of race is important because it is well established that older African Americans consistently perform at lower levels on cognitive function tasks compared to older Whites, and these differences often persist despite adjustments for education (e.g., Jones, 2003; Manly, 2005; Manly et al., 2002). While biological and social factors such as, physical illness or vascular disease (Crowe et al., 2010), educational quality (Crowe et al., 2013), and psychosocial constructs like perceived discrimination and stereotype threat (Barnes, Lewis, et al., 2012; Thames et al., 2013) are most commonly proposed as factors underlying differences in test performance, measurement *variance* is often offered as a possible explanation for the discrepancy in performance across race (Brickman, Cabo, & Manly, 2006; Jones, 2003; Manly, 2005). Our results demonstrate that measurement variance is not a reason for lower cognitive performance among the African Americans in this study. However, we did not test for other sources of bias (e.g., education or age); our focus on "strong-plus" rather than "strict" invariance as our highest level was in part an effort to minimize contributions from these sources, including any inherent test bias associated with the cognitive measures used in our battery. The findings that the

https://doi.org/10.1017/S1355617715001113 Published online by Cambridge University Press

group mean factor scores were lower, albeit not significantly so, for African Americans relative to Whites are consistent with the idea that differences in test performance between African Americans and Whites likely reflect a combination of multiple social and environmental experiences over the life course (e.g., Jones, 2003). For example, differences in literacy between older African Americans and Whites born before the historic Brown *versus* Board of Education decision are well established, and have been shown to not only attenuate racial differences in cognitive test performance (Manly et al., 2002), but also to be an important construct to account for differences compared to years of education or race-stratified normative data (Silverberg, Hanks, & Tompkins, 2013). Other race-relevant variables such as perceived discrimination (Barnes, Lewis, et al., 2012) and segregation (Aiken-Morgan, Gamaldo, Sims, Allaire, & Whitfield, 2014) have been shown to be important covariates of cognitive performance in African Americans as well. Older African Americans and Whites in the United States differ with respect to social environment, history, experience, cultural norms, beliefs, and attitudes. Future studies are needed to determine whether these differences may influence cognitive test performance, and whether these factors can account for the lower performance observed in this and other studies. Importantly, given that the underlying latent construct of our cognitive battery was found to be invariant across race, we are now free to investigate lifestyle factors and other racial experiences that may explain the lower performance.

The study has some limitations. First, all models considered in these analyses were *linear*. This does not preclude future tests of hypothesized *nonlinear* relationships among these cognitive test scores or among the factors conceptualized as representing "cognitive function." However, our linear modeling is consistent with current models of Alzheimer's disease and change in symptoms or biomarkers, which are typically derived as simple differences over time, that is, focused on linear change. While not necessarily ideal, linear models are simple to interpret and can bring statistical models closer to clinical applications, supporting this common modeling stance. Our model involves 18 tests and specialized estimation, and while a less complex conceptualization of "cognitive function" might be simpler to interpret, reducing the complexity of the cognitive battery might very well reduce the validity of the assessment overall—and might not be invariant as ours appears to be. Along this line, this study focused specifically on the level of invariance, and not on the model itself (which has been validated elsewhere; Wilson et al., 2004). Therefore, overlap in factor constructs or factor makeup was not studied and was instead treated as "given."

Second, our methodology was to start with the most constrained ("strong-plus") invariance and proceed through "strong" to "weak" invariance by iteratively releasing constraints for the models consistent with each level of invariance. An alternative approach is to begin with a freely estimated model and add constraints (moving from "none" to "weak" to "strong" to "strict" invariance) instead. This is more common and *is* advisable in projects where the measurement model is not well established, especially if the model fit has never been replicated in an independent sample. However, there might be highly localized invariant parameters within this model that were missed in our approach. Our argument is that, since cognition is so complex that the particular components of the model are unlikely to be meaningful measurement models on their own, this overall model fits the construct (cognitive function) sufficiently well and invariantly over time and race to yield useful and clinically meaningful assessments and—most importantly—comparisons over time. Because the measurement model we tested here has been used/fit in independent samples, we believed that the assumption that meeting our identified model fit criteria at any of the three levels would be sufficient to rule out any non-invariance that has practical or clinical significance.

Third, because MARS (from which the African American sample we analyzed was derived) started approximately 6 years after ROS and MAP, we were only able to model 2 years of follow-up data. Future studies with these cohorts will be able to include more years of longitudinal data collection. These analyses establish the invariance of our measurement model, so we are now more confident about pursuing structural models of cognitive change, and over longer time periods.

A final consideration is the variation in the strengths of associations of each test with its factor (factor loadings). Some factors have many fewer indicator tests than others, which is more of a function of the availability of valid cognitive tests than it is of our modeling or results. Although the results support the use and interpretability of factor scores that are derived from this model and its constituent tests, these results do not suggest that these assessments are "ideal" for capturing these cognitive domains. If other tests are preferred or used, these results suggest that establishing measurement invariance over time *and* race is achievable; best practices in the assessment and measurement fields dictate that they are also necessary.

This study also has strengths. First, our data come from three well-characterized epidemiologic cohorts with large sample sizes for both African Americans and Whites. Second, the five-factor cognitive model used in this study is fairly complex and factor analytic support for the five domains has been previously reported (Wilson et al., 2002). Third, the modeling approach that we used is consistent with best practices for studying and establishing measurement invariance (see e.g., Bollen & Curran, 2006; Horn & McArdle, 1992; Meredith & Teresi, 2006). Finally, our three cohorts were recruited with similar strategies, and cognitive function was assessed using a uniform and fairly elaborate collection of assessments. The cohorts do not represent a random sample from the general population of Whites or African Americans of this age range, and the assessments—while standard neuropsychological instruments—are not typically all assessed in the same individuals. Therefore, our cohorts all have similar years of education, and possibly other social similarities that we did not assess, and our model—and

its invariance—could have been affected by these factors. As stated previously, most racial differences stem from multiple cultural and social attributes often associated with race including, but not limited to, socioeconomic status, low literacy and education, racial discrimination, residential and school segregation, and lack of access to quality healthcare. Future studies will need to examine these and other factors in studies that compare cognitive performance between African Americans and Whites. However, it is important to note that the purpose of a measurement model is to represent whatever is being measured for anyone in whom it would be measured, and these results suggest that our complex model (Figure 1) for cognitive assessment functions as intended over time and across race (White and African American). Although there is clearly room for improvements, including stronger associations between indicators and factors, more indicators for some factors (e.g., semantic and working memories) and fewer for others (episodic memory), this validated measurement-invariant model of cognitive function in older African Americans and Whites can now be confidently used in other studies, including exploration of cognitive effects of neuropathology and studies of neuroimaging and other biomarkers.

ACKNOWLEDGMENTS

The authors thank the participants of the Minority Aging Research Study, the Rush Memory and Aging Project, and the Religious Order Study, for their invaluable contributions. We thank Charlene Gamboa, MPH; Tracy Colvin, MPH; Tracey Nowakowski, Barbara Eubeler, and Karen Lowe-Graham, MS, for study recruitment and coordination, and John Gibbons, MS and Greg Klein for data management, and the staff of the Rush Alzheimer's Disease Center. This research was supported by National Institute on Aging Grants (R01AG22018, R01AG17917, P3010161), and the Illinois Department of Public Health. The authors have no conflicts of interest.

REFERENCES

Aiken-Morgan, A.T., Gamaldo, A.A., Sims, R.C., Allaire, J.C., & Whitfield, K.E. (2014). Education desegregation and cognitive change in African American older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *70*, 348–356.

Albert, M., Smith, L.A., Scherr, P.A., Taylor, J.O., Evans, D.A., & Funkenstein, H.H. (1991). Use of brief cognitive tests to identify individuals in the community with clinically diagnosed Alzheimer's disease. *International Journal of Neuroscience*, *57*, 167–178.

Allison, P.D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557.

Arnold, B.R., Montgomery, G.T., Castaneda, I., & Longoria, R. (1994). Acculturation and performance of Hispanics on selected Halstead-Reitan neuropsychological tests. *Assessment*, *1*, 239–248.

Barnes, L.L., Lewis, T.T., Begeny, C.T., Yu, L., Bennett, D.A., & Wilson, R.S. (2012). Perceived discrimination and cognition in older African Americans. *Journal of the International Neuropsychological Society*, *18*, 856–865.

Barnes, L.L., Shah, R.C., Aggarwal, N.T., Bennett, D.A., & Schneider, J.A. (2012). The Minority Aging Research Study: Ongoing efforts to obtain brain donation in African Americans without dementia. *Current Alzheimer Research*, *9*, 734–745.

Bennett, D.A., Schneider, J.A., Arvanitakis, Z., & Wilson, R.S. (2012). Overview and findings from the Religious Orders Study. *Current Alzheimer Research*, *9*, 628–645.

Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., & Wilson, R.S. (2012). Overview and findings from the rush Memory and Aging Project. *Current Alzheimer Research*, *9*, 646–663.

Benton, A.L., Sivan, A.B., Hamsher, K.D., Varney, N.R., & Spreen, O. (1994). *Contributions to neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Blankson, A.N., & McArdle, J.J. (2013). Measurement invariances of cognitive abilities across ethnicity, gender, and time among older Americans. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *70*, 386–397.

Bollen, K.A., & Curran, P.J. (2006). *Latent curve models: A structural equation perspective*. New York, NY: Wiley.

Bowden, S.C., Weiss, L.G., Holdnack, J.A., & Lloyd, D. (2006). Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale–III. *Psychological Assessment*, *18*, 334–339.

Brewster, P.W.H., Melrose, R.J., Marquine, M.J., Johnson, J.K., Napoles, A., MacKay-Brandt, A., … Mungas, D. (2014). Life experience and demographic influences on cognitive function in older adults. *Neuropsychology*, *28*, 846–858.

Brickman, A.M., Cabo, R., & Manly, J.J. (2006). Ethical issues in cross-cultural neuropsychology. *Applied Neuropsychology*, *13*, 91–100.

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Byrne, B.M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B.M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Census Bureau. Retrieved from www.census.gov/population/www/socdemo.race.html

Chen, F.F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504.

Chen, F.F., Sousa, K.H., & West, S.G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471–492.

Cooper, J.A., & Sagar, H.J. (1993). Incidental and intentional recall in Parkinson's disease: An account based on diminished attentional resources. *Journal of Clinical and Experimental Neuropsychology*, *15*, 713–731.

Crowe, M., Clay, O.J., Martin, R.C., Howard, V.J., Wadley, V.G., Sawyer, P., & Allman, R.M. (2013). Indicators of childhood quality of education in relation to cognitive function in older adulthood. *The Journals of Gerontology: Series A: Biological Sciences and Medical Sciences*, *68*, 198–204.

Crowe, M., Sartori, A., Clay, O.J., Wadley, V.G., Andel, R., Wang, H.X., … Allman, R.M. (2010). Diabetes and cognitive decline: Investigating the potential influence of factors related to health disparities. *Journal of Aging and Health*, *22*, 292–306.

Early, D.R., Widaman, K.F., Harvey, D., Beckett, L., Park, L.Q., Farias, S.T., … Mungas, D. (2013). Demographic predictors of

cognitive change in ethnically diverse older persons. *Psychology & Aging*, *28*, 633–645.

Ekstrom, R.B., French, J.W., Harman, H.H., & Kermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Hayden, K.M., Jones, R.N., Zimmer, C., Plassman, B.L., Browndyke, J.N., Pieper, C., … Welsh-Bohmer, K.A. (2011). Factor structure of the National Alzheimer's Coordinating Centers uniform dataset neuropsychological battery: An evaluation of invariance between and within groups over time. *Alzheimer Disease and Associated Disorders*, *25*, 128–137.

Hertzog, C., & Schaie, K.W. (1986). Stability and change in adult intelligence: Analysis of longitudinal covariance structures. *Psychology and Aging*, *1*, 159–171.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jones, R.N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health*, *7*, 83–102.

Kaplan, E.F., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Philadelphia: Lea & Febiger.

Loewenstein, D.A., Arguelles, T., Arguelles, S., & Linn-Fuentes, P. (1994). Potential cultural bias in the neuropsychological assessment of the older adult. *Journal of Clinical and Experimental Neuropsychology*, *16*, 623–629.

Maitland, S.B., Intrieri, R.C., Schaie, K.W., & Willis, S.L. (2000). Gender differences and changes in cognitive abilities across the adult life span. *Aging Neuropsychology & Cognition*, *7*, 32–53.

Manly, J.J. (2005). Advantages and disadvantages of separate norms for African Americans. *The Clinical Neuropsychologist*, *19*, 270–275.

Manly, J.J., Jacobs, D.M., Touradji, P., Small, S.A., & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*, *8*, 341–348.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E.M. (1984). 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*, *34*, 939–944.

Meredith, W., & Teresi, J.A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69–S77.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Morris, J.C., Heyman, A., Mohs, R.C., Hughes, J.P., van Belle, G., Fillenbaum, G., … Clark, C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, *39*, 1159–1165.

Mulaik, S.A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Mungas, D., Widaman, K.F., Reed, B.R., & Tomaszewski, F.S. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology*, *25*, 260–269.

Muthén, L.K., & Muthén, B.O. (1998-2012). *Mplus User's Guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Raven, J.C., Court, J.H., & Raven, J. (1992). *Manual for Raven's progressive matrices and vocabulary*. Oxford: Oxford Psychologists Press.

Schaie, K.W., Willis, S.L., Jay, G., & Chipuer, H. (1989). Structural invariance of cognitive abilities across the adult life span: A cross-sectional study. *Developmental Psychology*, *25*, 652–662.

Schwartz, B.S., Glass, T.A., Bolla, K.I., Stewart, W.F., Glass, G., Rasmussen, M., … Bandeen-Roche, K. (2004). Disparities in cognitive functioning by race/ethnicity in the Baltimore Memory Study. *Environmental Health Perspectives*, *112*, 314–320.

Siedlecki, K.L., Honig, L.S., & Stern, Y. (2008). Exploring the structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. *Neuropsychology*, *22*, 400–411.

Siedlecki, K.L., Manly, J.J., Brickman, A.M., Schupf, N., Tang, M.X., & Stern, Y. (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology*, *24*, 402–411.

Silverberg, N.D., Hanks, R.A., & Tompkins, S.C. (2013). Education quality, reading recognition, and racial differences in the neuropsychological outcome from traumatic brain injury. *Archives of Clinical Neuropsychology*, *28*, 485–491.

Sloan, F.A., & Wang, J. (2005). Disparities among older adults in measures of cognitive function by race or ethnicity. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*, 242–250.

Smith, A. (1982). *Symbol Digit Modalities Test Manual - Revised*. Los Angeles: Western Psychological Services.

Thames, A.D., Hinkin, C.H., Byrd, D.A., Bilder, R.M., Duff, K.H., Mindt, M.R., … Streiff, V. (2013). Effects of stereotype threat, perceived discrimination, and examiner race on neuropsychological performance: Simple as black and white? *Journal of the International Neuropsychological Society*, *19*, 583–593.

Tractenberg, R.E., Aisen, P.S., Weiner, M.F., Cummings, J.L., & Hancock, G.R. (2006). Independent contributions of neural and 'higher order' deficits to symptoms in AD: A latent variable approach. *Alzheimer's & Dementia*, *2*, 303–313.

Trenerry, M.R., Crosson, B., DeBoe, J., & Leber, W.R. (1989). *Stroop Neuropsychological Screening Test Manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Tuokko, H.A., Chou, P.H., Bowden, S.C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *Journal of the International Neuropsychological Society*, *15*, 416–425.

Wechsler, D. (1987). *Wechsler Memory Scale-Revised manual*. San Antonio, TX: Psychological Corporation.

Wilson, R.S., Barnes, L.L., Krueger, K.R., Hoganson, G., Bienias, J.L., & Bennett, D.A. (2005). Early and late life cognitive activity and cognitive systems in old age. *Journal of the International Neuropsychological Society*, *11*, 400–407.

Wilson, R.S., Beckett, L.A., Barnes, L.L., Schneider, J.A., Bach, J., Evans, D.A., & Bennett, D.A. (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychology and Aging*, *17*, 179–193.

Wilson, R.S., Fleischman, D.A., Myers, R.A., Bennett, D.A., Bienias, J.L., Gilley, D.W., & Evans, D.A. (2004). Premorbid proneness to distress and episodic memory impairment in Alzheimer's disease. *Journal of Neurology, Neurosurgery, & Psychiatry*, *75*, 191–195.