Journal of the International Neuropsychological Society (2015), **21**, 506–518.

doi:10.1017/S1355617715000508

Predictors of Retest Effects in a Longitudinal Study of Cognitive Aging in a Diverse Community-Based Sample

Alden L. Gross,[1] Andreana Benitez,[2] Regina Shih,[3] Katherine J. Bangen,[4] M. Maria M. Glymour,[5] Bonnie Sachs,[6] Shannon Sisco,[7] Jeannine Skinner,[8] Brooke C. Schneider,[9] AND Jennifer J. Manly[10]

[1]Departments of Epidemiology and Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland; and Center on Aging and Health, Johns Hopkins University, Baltimore, Maryland
[2]Department of Radiology and Radiological Sciences, Center for Biomedical Imaging, Medical University of South Carolina, Charleston, South Carolina
[3]RAND Corporation, Arlington, Virginia
[4]Department of Psychiatry, University of California, San Diego, La Jolla, California
[5]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California
[6]Department of Neurology, Wake Forest Baptist Medical Center, Winston Salem, North Carolina
[7]North Florida/South Georgia Veterans Health System, Department of Veterans Affairs, Gainesville, Florida
[8]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, Washington
[9]Department of Psychiatry and Psychotherapie, University Hospital Hamburg-Eppendorf, Hamburg, Germany
[10]Taub Institute for Research on Alzheimer's disease and the Aging Brain, Columbia University, New York, New York; Gertrude H. Sergievsky Center, Columbia University, New York, New York; and Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, New York

(RECEIVED November 12, 2014; FINAL REVISION May 14, 2015; ACCEPTED May 26, 2015)

Abstract

Better performance due to repeated testing can bias long-term trajectories of cognitive aging and correlates of change. We examined whether retest effects differ as a function of individual differences pertinent to cognitive aging: race/ethnicity, age, sex, language, years of education, literacy, and dementia risk factors including apolipoprotein E ε4 status, baseline cognitive performance, and cardiovascular risk. We used data from the Washington Heights-Inwood Columbia Aging Project, a community-based cohort of older adults ($n = 4073$). We modeled cognitive change and retest effects in summary factors for general cognitive performance, memory, executive functioning, and language using multilevel models. Retest effects were parameterized in two ways, as improvement between the first and subsequent testings, and as the square root of the number of prior testings. We evaluated whether the retest effect differed by individual characteristics. The mean retest effect for general cognitive performance was 0.60 standard deviations (95% confidence interval [0.46, 0.74]), and was similar for memory, executive functioning, and language. Retest effects were greater for participants in the lowest quartile of cognitive performance (many of whom met criteria for dementia based on a study algorithm), consistent with regression to the mean. Retest did not differ by other characteristics. Retest effects are large in this community-based sample, but do not vary by demographic or dementia-related characteristics. Differential retest effects may not limit the generalizability of inferences across different groups in longitudinal research. (*JINS*, 2015, *21*, 506–518)

Keywords: Practice effect, Retest effect, Neuropsychological testing, Older adults, Gerontology, Individual differences

INTRODUCTION

Estimation of the pace of cognitive decline throughout the lifecourse is central to research on cognitive aging and dementia (Salthouse, 2010a). Cognitive decline is a more compelling marker of Alzheimer's disease (AD) dementia than impairment at one testing session because it is less affected by historical factors such as years of education that precede the onset of AD (Glymour et al., 2005). However, design and analysis of longitudinal studies, wherein cognitive testing is repeatedly conducted on the same person over time, can be complicated because, in addition to normal aging or maturation, factors such as selective attrition, period and cohort effects, statistical artifacts (e.g., regression to the mean), and retest or practice effects contribute to changes in cognitive test performance (Dodge et al., 2011; Salthouse, 2010a, 2010b).

Retest or practice effects refer to the extent to which repeated cognitive testing results in improved performance due to familiarity with the testing materials and setting (Horton, 1992; Zehnder, Blasi, Berres, Spiegel, & Monsch, 2007). These effects are well-documented in longitudinal studies of

Correspondence and reprint requests to: Alden L. Gross, Johns Hopkins Bloomberg School of Public Health, 2024 E. Monument St., Baltimore, MD 21205, USA. E-mail: agross14@jhu.edu

cognitive aging (Abner et al., 2012; Basso, Bornstein, & Lang, 1999; Calamia, Markon, & Tranel, 2012; Collie, Maruff, Darby, & McStephen, 2003; Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004; Duff et al., 2011; Ferrer, Salthouse, Stewart, & Schwartz, 2004; Ferrer, Salthouse, McArdle, Stewart, & Schwartz, 2005; Frank, Wiederholt, Kritz-Silverstein, Salmon & Barrett-Connor, 1996; Horton, 1992; Howieson et al., 2008; Ivnik et al., 1999; Jacqmin-Gadda, Fabrigoule, Commenges, & Dartigues, 1997; Machulda et al., 2013; Mitrushina and Satz, 1991; Rabbitt, Diggle, Smith, Holland, & McInnes, 2001; Rabbitt, Diggle, Holland, & McInnes 2004; Salthouse, 2009; Wilson, Leurgans, Boyle, & Bennett, 2011; Wilson, Li, Bienias, & Bennett, 2006; Zehnder et al., 2007). Although strongest with shorter retest intervals, retest effects have been documented for up to 5 (Burke, 1997) and 12 (Salthouse, Schroeder, & Ferrer, 2004) years (Hausknecht, Halpert, Di Paolo, & Gerrard, 2007). A consensus conference for clinical neuropsychology has called for research on ramifications of repeated cognitive testing (Heilbronner et al., 2010). Van der Elst, Van Boxtel, Van Breukelen, and Jolles (2008) found a robust increase of between 0.2 and 0.6 standard deviations (*SD*) in verbal list-learning performance 3 years after the first testing occasion in a large sample of cognitively normal older adults, while Bartels, Wegrzyn, Wiedl, Ackermann, and Ehrenreich (2010) found medium to large retest effects between 0.36 and 1.19 *SD* after approximately 3 months. Although both of these studies conceptualize retest effects as a one-time boost between the first and subsequent occasions, retest effects may also exist at each visit with diminishing returns (Collie et al., 2003; Sliwinski, Hoffman, & Hofer, 2010).

In epidemiologic research, failure to account for retest effects obscures the estimated rate of cognitive decline. If retest effects are correlated with risk factors of interest, ignoring them may lead to biased estimates of their effects on the rate of cognitive change. Retest effects may differ by the type of cognitive task. Tests that measure different cognitive abilities (e.g., memory, language) (Cooper et al., 2004) or that use different administration or response modalities (e.g., oral *vs*. written) might show different patterns of retest effects. In this study, we examined retest effects at the level of constructs rather than individual cognitive tests to avoid detecting differences in modality.

In addition to the type of test, retest effects may be attributable to participant characteristics related to proficiency in test-taking *via* test-taking strategies and less test anxiety, in which case persons with less testing experience might show larger retest effects (Thorndike, 1922). Retest effects may also be attributed to episodic memory, or the successful learning and retention of test content such that subsequent improved performance is facilitated by recollection of the content. This is a motivation behind the use of alternate forms for tests of episodic memory (e.g., Benedict and Zgaljardic, 1998; Delis, Kramer, Kaplan, & Ober, 2000). Thus, testing for differential retest effects by factors related to test experience and episodic memory provide a way to better understand retest effects.

Clinically, group-level differences in retest effects have implications for test–retest reliability and interpretation of norms. The rank ordering of patients at one assessment compared to another may be stable despite a large retest effect, indicating good test–retest reliability but complicating interpretations from a single set of norms for both assessments (Calamia et al., 2012). This would interfere with tracking of disease progression and detection of decline. Alternatively, if test–retest reliability is moderate or low in an overall sample but higher in subgroups, that could reflect systematic group differences in the magnitude of individual differences in the amount of retest (Salthouse & Tucker-Drob, 2008).

Sociodemographic Factors Related to Test Experience

Because educational attainment is a strong predictor of cognitive performance in later life, retest effects may differ by number of years of education (Cagney & Lauderdale, 2002; Stern et al., 1994; Stuss, Stethem, & Poirier, 1987). Individuals with less education or lower quality education have less prior experience with test-taking and strategies for maximizing test performance. Such individuals have the most to gain from practice with the test. Similarly, given differences in early educational experiences for older adults by race and ethnicity due to persistent educational inequalities (Glymour & Manly, 2008), we hypothesized that Hispanic older adults, most of whom in the present sample are immigrants to the United States, may be less familiar on average with testing and therefore experience greater retest effects (Gould, 1996).

Age, sex, and language spoken at home may also moderate retest effects. Previous research suggests that, with the exception of measures of word list recall, retest effects are inversely related to age (Mitrushina & Satz, 1991; Rabbitt, Lunn, Wong, & Cobain, 2008). Sex differences in cognitive performance have been documented for a range of cognitive abilities, suggesting differential retest effects may also occur. Women tend to do better on memory tests (leaving men with more room to improve upon retest), while men tend to do better on visuospatial tasks (Mann, Sasanuma, Sakuma, & Masaki, 1990; Salthouse, 2010a; Voyer, Voyer, & Bryden, 1995). Primary language may also be important to retest; one study found Spanish-speakers demonstrated greater retest effects than English-speakers (Mungas, Reed, Marshall, & Gonzalez, 2000).

Dementia Risk Factors

The ability to learn and retain new information may also facilitate retest effects. Previous studies suggest that the absence of retest effects may reflect amnestic mild cognitive impairment (MCI) or AD (Darby, Maruff, Collie, & McStephen, 2002; Duff et al., 2011; Frank et al., 1996; Schrijnemaekers, de Jager, Hogervorst, & Budge, 2006). However, at least one recent study reported retest effects for memory in participants with MCI and dementia (Machulda et al., 2013). The apolipoprotein E (APOE) ε4 allele predicts

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

earlier onset of Alzheimer's disease among older Whites (Baxter, Caselli, Johnson, Reiman, & Osborne, 2003; Blair et al., 2005; Haan, Shemanski, Jagust, Manolio, & Kuller, 1999), but the association seems to be attenuated among Blacks (Borenstein, Copenhaver, & Mortimer, 2006; Tang et al., 1998). A previous study found APOE e4 carriers did not exhibit a retest effect (Machulda et al., 2013). Furthermore, cardiovascular burden is an established risk factor for poorer cognition and neurodegenerative disease, especially among minority older adults (Flicker, 2010; Luchsinger et al., 2005). Thus, it is possible that greater cardiovascular risk burden may affect the magnitude of retest effects.

The Present Study

We examined whether retest effects vary by demographic factors such as race/ethnicity, age, language spoken at home, literacy, sex, years of education, and dementia risk factors including APOE ε4 status, baseline cognitive status, and cardiovascular burden. We estimated multilevel random effects models of change in general cognitive performance, memory, executive function, and language. The mean retest effect was allowed to differ by the characteristic of interest. We hypothesized that Hispanic racial/ethnic group membership and fewer years of education predict larger retest effects, while dementia risk factors such as possession of the APOE e4 allele, lower cognitive performance at baseline, and greater cerebrovascular risk burden predict smaller retest effects.

METHODS

Participants and Procedures

We used data on $N = 4073$ participants from the Washington Heights-Inwood Columbia Aging Project (WHICAP), an ongoing epidemiologic cohort of community-living Medicare-eligible older adults recruited from northern Manhattan (Tang et al., 2001). Participants were residents of three contiguous US census tracts in Northern Manhattan, New York. Individuals were invited to participate in an in-person survey in 1992, with follow-up visits every 2 to 3 years. Recruitment re-opened in 1999 to replenish the cohort. At each interview, participants answered extensive questionnaires about their early life education, health, and cognitive performance. The present study used data from 4073 participants who participated in neuropsychological assessments. Details of the sampling strategies and recruitment outcomes have been published previously (Luchsinger, Tang, Shea, & Mayeux, 2001; Manly, Schupf, Tang, & Stern, 2005). The study was approved by Institutional Review Boards at Columbia Presbyterian Medical Center, Columbia University Health Sciences, and the New York State Psychiatric Institute.

Measures

Racial/ethnic group

Participants self-reported their race by selecting membership from categories of American Indian/Alaska Native, Asian, Native Hawaiian or other Pacific Islander, Black or African American, or White. Participants were then asked whether they were Hispanic. We grouped participants into categories of non-Hispanic White, non-Hispanic Black, and Hispanic.

Cardiovascular burden

We used a summary of cardiovascular burden based on presence of diabetes, hypertension, heart disease, stroke, central obesity, and current smoking status (Schneider et al., 2014).

Educational experience

We used self-reported years of education completed to represent previous exposure to learning. However, there is considerable heterogeneity in the amount of learning obtained given a certain grade level due to interstate, racial, and international differences in educational quality (Glymour & Manly, 2008; Manly et al., 1999, 2002, 2003, 2004). Thus, years of education is a poor proxy for schooling. Because of this, we also tested for differences in retest by level of literacy as a proxy for quality of education (Manly et al., 2004). We stratified this analysis by language of administration due to nonequivalence of the English WRAT (Wilkinson & Robertson, 2006) and Spanish WAT (Del Ser, Gonzalez-Montalvo, MartinezEspinosa, Delgado-Villapalos, & Bermejo, 1997). We took a median split of performance on these tests.

Cognitive performance

WHICAP administered a neuropsychological test battery at each study visit (Tang et al., 2001). Tests were designed for administration in Spanish or English (Dugbartey, Townes, & Mahurin, 2000; Jacobs et al., 1995). The tests are described in the Appendix. We constructed factor scores for general cognitive performance, memory, executive function, and language using confirmatory factor analysis models for each domain. We used immediate recall, delayed recall, and delayed recognition from the Buschke Selective Reminding Test to construct the memory factor. The executive functioning factor was derived using the Color Trail-Making Test (A and B), WAIS Similarities, Identities/Oddities, shape time, time to detect a consonant trigram, phonemic fluency, and semantic fluency for animals. Language was derived using phonemic and semantic fluency, 15-item Boston Naming, repetition, and comprehension. All of the above variables contributed to the general cognitive factor. The assignment of tests to factors is largely consistent with a previously published factor analysis of the neuropsychological test battery in the WHICAP cohort (Siedlecki et al., 2010), except that we dropped the speed factor derived by Siedlecki et al. and added an executive functioning factor. The executive functioning factor has more indicators that represent broader fluid ability, and is more reliable than two separate factors for speed and reasoning that are each based on fewer measures. Previous factor analysis of the WHICAP battery revealed that semantic and phonemic fluency both load best with language (Siedlecki et al., 2010), which is also consistent with the derived

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

executive functioning factor in the Alzheimer's Disease Neuroimaging Initiative (Gibbons et al., 2012).

Each factor was scaled to have a mean of 50 and standard deviation (*SD*) of 10 in the US population of adults aged 70 years and older to facilitate comparison of magnitudes of effects across domains and with future studies. Details are provided elsewhere (Gross, Jones, Fong, Tommet, & Inouye, 2014; Gross, Sherva, et al., 2014). Briefly, we calibrated the factors using a nationally representative sample of adults aged 70 and older from the Aging, Demographics, and Memory Study (ADAMS), a sub-study of the Health and Retirement Study (Juster & Suzman, 1995; Langa et al., 2005). The ADAMS battery included Trails A and B, Digits Forward and Backward, semantic and phonemic fluency, Boston Naming Test, Symbol Digit Modalities, and a 10-noun word recall task. Items common to ADAMS and WHICAP served as links to calibrate cognitive factors. The factor analysis was performed in a longitudinal dataset with multiple records per participant. We fixed item discrimination and difficulty parameters for common items in the factor analysis including both WHICAP and ADAMS to the values estimated in an ADAMS-only factor analysis. This scaling approach does not make the WHICAP sample nationally representative, but it allows future analysts, using other datasets with items overlapping with ADAMS, to derive directly comparable scores. The approach assumes measurement invariance of factors with respect to time: an assumption previously verified in other samples of older adults and which we evaluated in WHICAP through formal tests described earlier (Hayden et al., 2011; Johnson et al., 2012). We additionally tested longitudinal measurement invariance of the factors among participants assessed at baseline and whose second study visit was between 1.5 and 2.5 years later (median: 2.1 years) using multiple group confirmatory factor analysis models. Details are provided in the Appendix.

Analyses

To test hypotheses, we used multilevel models with random effects for people and time alongside fixed effects for retest in general cognition, memory, executive functioning, and language (Johnson et al., 2012; Laird & Ware, 1982; Muthén & Curran, 1997; Raudenbush & Bryk, 2002). Time since enrollment into the study was the timescale of interest. The system of equations below describes the basic model:

$$\text{Level 1}: Y_{ij} = \beta_{0i} + \beta_{1i} * \text{time}_{ij} + \beta_2 * \text{retest}_{ij} + \sum_{p=3}^{5} \beta_p * X_i + \varepsilon_{ij} \quad (1)$$

$$\text{Level 2}: \beta_{0i} = \gamma_{00} + U_{0i} \quad (2)$$

$$\beta_{1i} = \gamma_{10} + U_{1i} \quad (3)$$

Y_{ij} is a cognitive outcome (general cognitive performance, memory, executive functioning, or language) for participant i at time j. The level 1 model describes within-person change over time based on random (U_{0i}) and fixed (γ_{00}) effects for participants, random (U_{1i}) and fixed (γ_{10}) effects for time, a fixed effect (β_2) for the retest effect, adjustment variables β_p, and residual error ε_{ij} for each participant at each time. Level 2 equations describe the random and fixed effects for participants and time. Distributions of ε_{ij}, U_{0i}, and U_{1i} are assumed to be normal with mean 0 and variance 1.

We coded retest in two ways to acknowledge different conceptualizations of how they come about. First, as our primary analysis, the retest variable was coded 0 at each participant's first study visit in which they were administered a neuropsychological battery, and 1 otherwise. Retest effects here are interpretable as the difference or jump in performance from the first assessment to the predicted performance based on the level and slope of change at the second and later assessments. This characterization is consistent with previous studies examining retest effects (Ivnik et al., 1999; Rabbitt et al., 2004; Salthouse et al., 2004; Salthouse & Tucker-Drob, 2008). Previous studies have suggested that subsequent gains after the second testing occasion are negligible (but see the Discussion section) (Kausler, 1994; Rabbitt, 1993, 2004). Second, to allow for the possibility that participants learn more at each test occasion with diminishing returns over time (Abner et al., 2012; Collie et al., 2003; Hoffman et al., 2011), we also calculated retest as the square root of the number of prior test occasions. We adjusted all models for sex, baseline age, and recruitment cohort (1992 or 1999).

To determine whether retest effects vary by individual characteristics, or effect modification, we extended the model described above to a series of multiple group models in a structural equation modeling framework, in which groups were defined based on the characteristic of interest. Groups were defined by race/ethnicity (non-Hispanic Black, non-Hispanic White, Hispanic), age (<75 years, 75–80 years, 80+ years), sex, years of education (less than 8 years, 8 or more years), literacy (median split, models conducted separately by language of administration), APOE ε4 status (carrier, noncarrier), quartile of baseline general cognitive performance, and number of cardiovascular risk factors (0, 1, 2, or ≥3). We conducted analyses by baseline quartiles of cognitive performance instead of adjudicated dementia diagnosis because, in WHICAP, neuropsychological test performance was considered during the adjudication procedure. Nonetheless, we also conducted a sensitivity analysis to identify whether excluding participants with an adjudicated diagnosis of dementia affected results. Differences in mean retest effects between these groupings are estimated in a manner analogous to using an interaction between the characteristic and retest indicator, as follows:

$$\text{Level 1}: Y_{ij} = \beta_{0i} + \beta_{1i} + \text{time}_{ij} + \beta_2 + \text{retest}_i + \beta_3 + \text{moderator}_i + \beta_4 + \left(\text{moderator}_i + \text{retest}_{ij}\right) + \varepsilon_{ij} \quad (4)$$

The interaction of the moderator and the retest effect, β_4, is the parameter of interest. In planned sensitivity analyses, we examined retest effects for all component tests in the WHICAP neuropsychological battery. Analyses were conducted with Mplus statistical software (version 7.11, Muthén & Muthén, 1998–2012) using robust maximum likelihood

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

estimation that assumed outcome observations were missing at random, conditional on covariates (Little & Rubin, 1987). Fit of modeled trajectories to data was assessed with a pseudo-R^2 statistic. The pseudo-R^2 represents the proportion of variability in observed data explained by the model (Singer & Willet, 2003). It is calculated by squaring the correlation between observed and model-estimated (including random effects terms) outcome scores. We adjusted models for potential selective survival using inverse probability weights (Hernan & Robins, 2006) calculated from a logistic regression of death on age, sex, baseline general cognitive performance, APOE e4 status, education in years, recruitment cohort, and cardiovascular risk measured at baseline.

RESULTS

The study sample was mostly female (68.5%), had 8 or more years of education (53.6%), and the average age at the first visit was 77 years (range, 63–103 years) (Table 1). The sample was ethnically diverse, with 33.7% non-Hispanic Black, 24.9% non-Hispanic White, and 41.4% Hispanic. The percentage of participants with at least one APOE ε4 allele was 22.6%. Two-year test–retest reliabilities for the factors representing general cognitive performance, memory, executive functioning, and language were r = 0.88, r = 0.77, r = 0.80, and r = 0.83, respectively.

Overall Retest Effect

The median number of study visits was three (interquartile interval: 2, 4) and median follow-up time was 3.9 years (interquartile interval: 1, 7.8). The second study visit took place on average 1.9 years (interquartile interval: 1.3, 2.2 years) after the first study visit. For each cognitive outcome, a 1-point difference is analogous to a 0.1 standard deviation difference. As expected, the overall retest effect was considerable for all domains. For general cognitive performance, the retest effect was 0.60 *SD*, while the annual rate of general cognitive decline was only –0.047 *SD* (Table 2). Thus, the absolute value of the retest effect is the same magnitude as 12.8 years of cognitive decline (Table 2). Retest effects were also large for memory (retest = 0.57 *SD*; 95% confidence interval [CI] [0.42, 0.72] *SD*), executive functioning (retest = 0.45 *SD*; 95% CI [0.32, 0.58] *SD*), and language (retest = 0.64 *SD*; 95% CI [0.47, 0.81] *SD*) (Table 2).

Effect Modification of Retest Effects by Participant Characteristics

Models fit well to the data, with pseudo-R^2 values above 0.79 for each cognitive outcome (Table 3). Visual inspection of model residuals confirmed adequate fit to the data. The magnitude of the retest effect, parameterized as the jump from the first to subsequent test occasions, was statistically significant and positive for general cognitive performance in nearly every subgroup (Table 3). Inferences were similar for memory and for language. For executive functioning, average retest effects tended to be smaller but were mostly statistically significant (Table 3). This pattern of results was identical when we parameterized retest effects as the square root of the number of prior test occasions (Appendix Table 2).

The magnitude of retest effects did not differ significantly by race/ethnicity, age, language, sex, education, literacy, APOE status, or cardiovascular burden (Table 3). Participants in the lowest quartile of baseline general cognitive performance demonstrated greater retest effects compared to participants in the middle two quartiles of general cognitive performance, for whom retest effects were not significant (Table 3). Figure 1 shows the model-estimated cognitive

Table 1. Demographic characteristics: Results from WHICAP (*N* = 4073)

Variable	WHICAP sample (*N* = 4073)	Observed range
Age, mean (*SD*)	77.3 (7.0)	(63.0, 103.0)
Male, *n* (%)	1283 (31.5)	
Recruitment cohort, *n* (%)		
1992	1932 (47.4)	
1999	2141 (52.6)	
Years of follow-up, median (interquartile interval)	3.9	(1.0, 7.8)
Years between first and second testing, median (interquartile interval)	1.9	(1.3, 2.2)
Number of study visits, median (interquartile interval)	3.0	(2.0, 4.0)
Race/ethnicity, *n* (%)		
Non-Hispanic White	1013 (24.9)	
Non-Hispanic Black	1372 (33.7)	
Hispanic	1688 (41.4)	
Years of education, *n* (%)		
7 years or less	1883 (46.4)	
8 or more years	2179 (53.6)	
APOE ε4 status, *n* (%)		
Possesses ε4 allele	3151 (77.4)	
No ε4 allele	922 (22.6)	
Vascular risk factors, *n* (%)		
None	884 (21.7)	
One	1414 (34.7)	
Two	1167 (28.7)	
Three or more	608 (14.9)	
Cognitive factor scores, mean (*SD*)		
General cognitive performance	46.3 (12.0)	(–11.4, 87.5)
Memory	55.2 (13.4)	(13.8, 96.0)
Executive functioning	44.8 (11.0)	(11.7, 84.8)
Language	44.6 (10.7)	(5.7, 80.4)
Predominant language spoken at home, *n* English (%)	1626 (59.6)	
Quartile of general cognitive performance, *n* (%)		
First (lowest) quartile	1027 (25.2)	
Second quartile	1018 (25.0)	
Third quartile	1015 (24.9)	
Fourth (highest) quartile	1013 (24.9)	

SD = standard deviation.

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

Table 2. Retest effects and slopes for general and domain-specific cognitive performance: Results from WHICAP ($N = 4073$)

Cognitive outcome	Retest effect [95% CI]	Mean annual rate of decline [95% CI]	Ratio of retest to annual rate of decline
General cognitive performance	6.01 (4.58, 7.43)	–0.47 (– 0.48, –0.46)	12.78
Memory	5.72 (4.21, 7.22)	–0.54 (– 0.56, –0.52)	10.57
Executive functioning	4.46 (3.15, 5.78)	–0.42 (–0.44, –0.41)	10.55
Language	6.44 (4.74, 8.13)	–0.38 (–0.40, –0.36)	16.93

Note. Parallel process latent growth models of changes in global cognition, memory, and executive functioning score changes using time in study as the timescale. Each cognitive score was scaled to have a mean of 50 and standard deviation of 10 at the baseline study visit. The annual rate of decline is the mean of the random slope in the model. The ratio of retest and slope reflects the relative magnitude of the retest effect compared to subsequent annual cognitive decline. The retest parameters correspond to β_2 parameters in Eq. [1]. The model-estimated proportion of total variance attributable to between-persons differences was 86%, 74%, 78%, and 81% for general cognitive performance, memory, executive functioning, and language, respectively.

trajectory for participants at these quartiles of cognitive function. Although we did not exclude participants who had an adjudicated diagnosis of dementia in WHICAP, we observed that 645 of 679 (94.9%) of participants with dementia were in the lowest quartile of baseline cognitive performance (sensitivity), and 3006 of 3369 (89.2%) of non-demented participants had a score above the lowest quartile (specificity).

Sensitivity Analyses

We examined the magnitude of retest effects for each component test in the WHICAP battery. Results of this sensitivity analysis were consistent with findings using the factor scores. Retest effects were generally greater in magnitude for memory tests than for executive functioning tests. We also reran analyses excluding participants with dementia; the only change in inferences was that the difference in retest by baseline cognitive quartile was no longer statistically significant (Appendix 3). Although overall retest among participants with a study diagnosis of dementia did not statistically significantly differ from others in the lowest quartile of baseline general cognitive performance for any cognitive outcome ($p > .05$), participants with dementia did on average have higher retest effects for general cognitive performance (retest no dementia: –0.03 points, 95% CI [– 1.40, 1.34]; retest dementia: 2.96 points, 95% CI [1.65, 4.27]), memory (retest no dementia: –0.20 points, 95% CI [– 4.57, 4.17]; retest dementia: 4.33 points, 95% CI [2.31, 6.35]), executive functioning (retest no dementia: –0.31 points, 95% CI [– 2.11, 1.49]; retest dementia: 1.36 points, 95% CI [0.13, 2.59]), and language (retest no dementia: –0.36 points, 95% CI [– 2.44, 1.72]; retest dementia: 2.08 points, 95% CI [0.47, 3.69]).

DISCUSSION

In this large, diverse community-based sample of older adults, we examined differences in retest effects by racial/ethnic group, age, language spoken at home, sex, years of education, literacy, APOE ε4 status, baseline cognitive function, and cardiovascular burden. Despite the relatively long 2-year interval between testing occasions, the overall magnitude of retest was on average more than 10 times the annual rate of subsequent cognitive decline, and greatest for language. The magnitude of retest is in line with previous findings (Bartels et al., 2010; Van der Elst et al., 2008). The magnitude of retest effects did not differ by any characteristic examined other than baseline cognitive status: on average, participants performing in the lowest quartile at baseline experienced the greatest boost from repeated testing. This finding is probably attributable to regression to the mean. Overall, the results suggest retest effects do not differ greatly across observable demographic and dementia-related factors.

Previous research indicates that the magnitude of retest effects varies widely across different tests (Calamia et al., 2012; Frank et al., 1996), with effects typically but not always largest for visual memory and smallest for visuospatial ability (Calamia et al., 2012, but see also Dodrill & Troupin, 1975; Ferrer et al., 2004; Frank et al., 1996; McCaffrey, Onega, Orsillo, Nelles, & Haase, 1992). In our study, we built on prior research by considering cognitive domains instead of individual tests in an attempt to draw conclusions at the level of constructs, and mitigate the potential for spurious findings from multiple tests. A further implicit advantage of our study was the choice of scaling to an external standard, the ADAMS HRS. This scaling made no difference in the results compared to factors scores that were scaled internally. Scale choice is in many cases arbitrary. However, we believe that future scientific progress in the area of cognitive aging will be accelerated if findings are presented on a common scale across studies. Resources are available that describe how other studies can be linked to an external metric such as the nationally representative sample used here (e.g., Gross, Jones, et al., 2014; Gross, Sherva, et al., 2014; Jones et al., 2010).

Our data suggest retest effects were greater for participants in the lowest quartile of baseline cognitive performance. Sensitivity analyses revealed the largest practice effects were observed in the subgroup of participants diagnosed with dementia at baseline. Although in clinical settings, it is less likely to observe significant improvement in neuropsychological test performance in dementia patients upon follow-up, this phenomenon is not unusual in research settings. Participants who meet research criteria for dementia or MCI at do not always meet criteria at their next visit; this has been documented in WHICAP (Schofield et al., 1995; Manly et al., 2008) as well as in other population-based cohorts

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

Table 3. Retest effects for general and domain-specific cognitive performance: Results from WHICAP ($N = 4073$)

	General cognitive performance	Memory	Executive functioning	Language
Parameter	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Race/ethnicity				
Non-Hispanic White ($n = 1013$)	4.95* (1.87, 8.03)	5.81* (1.38, 10.24)	3.34* (0.71, 5.97)	5.65* (2.26, 9.04)
Non-Hispanic Black ($n = 1372$)	4.28* (1.69, 6.87)	3.85* (1.26, 6.44)	3.98* (1.61, 6.35)	3.92* (0.69, 7.15)
Hispanic ($n = 1688$)	3.49* (1.88, 5.10)	3.35* (1.66, 5.04)	2.19* (0.52, 3.86)	3.56* (1.54, 5.58)
Group differences				
White - Black	0.67 (–3.35, 4.69)	1.97 (–3.15, 7.09)	– 0.64 (–4.19, 2.91)	1.74 (–2.94, 6.42)
White - Hispanic	1.46 (–2.01, 4.93)	2.46 (–2.28, 7.20)	1.14 (–1.98, 4.26)	2.09 (–1.85, 6.03)
Black - Hispanic	0.79 (–2.27, 3.85)	0.50 (–2.58, 3.58)	1.78 (–1.12, 4.68)	0.36 (–3.46, 4.18)
Pseudo-R^2	0.90	0.80	0.81	0.86
Age				
Under 75 ($n = 1667$)	2.93* (0.09, 5.77)	2.84 (–0.02, 5.70)	2.64 (–0.05, 5.33)	3.39 (–0.24, 7.02)
Age 75 to 80 ($n = 1158$)	3.66* (1.31, 6.01)	3.59* (0.92, 6.26)	2.92* (0.65, 5.19)	3.36 (–0.15, 6.87)
80 and over ($n = 1248$)	4.02* (2.41, 5.63)	3.37* (1.78, 4.96)	2.10* (0.71, 3.49)	4.21* (2.43, 5.99)
Group differences				
Under 75 - (75 to 80)	– 0.73 (–4.41, 2.95)	– 0.74 (–4.64, 3.16)	– 0.28 (–3.81, 3.25)	0.03 (–5.03, 5.09)
Under 75 - Over 80	– 1.09 (–4.36, 2.18)	– 0.53 (–3.80, 2.74)	0.54 (–2.50, 3.58)	– 0.82 (–4.86, 3.22)
(75 to 80) - Over 80	– 0.36 (–3.22, 2.50)	0.22 (–2.88, 3.32)	0.82 (–1.85, 3.49)	– 0.85 (–4.79, 3.09)
Pseudo-R^2	0.90	0.80	0.81	0.86
Sex				
Male ($N = 1283$)	3.92* (1.41, 6.43)	3.45* (0.84, 6.06)	3.42* (1.15, 5.69)	4.65* (1.61, 7.69)
Female ($N = 2790$)	4.77* (3.01, 6.53)	4.80* (2.94, 6.66)	3.50* (1.78, 5.22)	4.72* (2.54, 6.90)
Group differences				
Female - Male	0.85 (–2.21, 3.91)	1.34 (–1.85, 4.53)	0.07 (–2.77, 2.91)	0.07 (–3.67, 3.81)
Pseudo-R^2	0.88	0.79	0.82	0.85
Predominant language spoken at home				
English ($n = 1572$)	3.69* (1.48, 5.90)	3.37* (1.10, 5.64)	2.51* (0.12, 4.90)	4.04* (1.14, 6.94)
Non-English ($N = 1009$)	5.69* (2.81, 8.57)	5.90* (2.61, 9.19)	4.88* (2.21, 7.55)	5.80* (2.31, 9.29)
Group differences				
English - Non-English	2.00 (–1.63, 5.63)	2.53 (–1.47, 6.53)	2.38 (–1.21, 5.97)	1.76 (–2.79, 6.31)
Pseudo-R^2	0.88	0.781	0.807	0.846
Years of education				
7 years or less ($n = 1883$)	3.41* (2.10, 4.72)	3.16* (1.81, 4.51)	2.30* (1.01, 3.59)	3.37* (1.78, 4.96)
8 or more years ($n = 2179$)	4.58* (2.76, 6.40)	4.76* (2.53, 6.99)	3.57* (1.85, 5.29)	4.71* (2.46, 6.96)
Group differences				
(8 or more years) - (7 years or less)	1.17 (–1.08, 3.42)	1.60 (–1.01, 4.21)	1.28 (–0.88, 3.44)	1.34 (–1.42, 4.10)
Pseudo-R^2	0.90	0.80	0.81	0.86
Literacy (Spanish speakers, $N = 746$)				
Lowest 50%	1.05 (–1.75, 3.85)	1.77 (–3.40, 6.94)	0.03 (–2.93, 2.99)	1.25 (–2.36, 4.86)
Highest 50%	1.92 (–4.10, 7.94)	2.22 (–6.35, 10.79)	1.43 (–6.45, 9.31)	2.54 (–7.06, 12.14)
Group differences				
(Highest) - (Lowest)	0.87 (–5.77, 7.51)	0.45 (–9.55, 10.45)	1.40 (–7.01, 9.81)	1.29 (–8.98, 11.56)
Pseudo-R^2	0.89	0.80	0.80	0.85
Literacy (English speakers, $N = 1569$)				
Lowest 50%	2.36 (–1.15, 5.87)	3.40 (–2.79, 9.59)	2.17 (–0.50, 4.84)	1.94 (–2.72, 6.60)
Highest 50%	2.61 (–1.82, 7.04)	3.86 (–7.65, 15.37)	2.09 (–0.36, 4.54)	2.83 (–0.84, 6.50)
Group differences				
(Highest) - (Lowest)	0.25 (–5.39, 5.89)	0.46 (–12.59, 13.51)	– 0.07 (–3.68, 3.54)	0.89 (–5.03, 6.81)
Pseudo-R^2	0.90	0.80	0.81	0.86
APOE ε4 status				
No ε4 allele ($n = 3151$)	4.57* (2.92, 6.22)	4.43* (2.67, 6.19)	3.49* (1.92, 5.06)	4.80* (2.84, 6.76)
Possesses ε4 allele ($n = 922$)	4.22* (1.20, 7.24)	4.23* (1.09, 7.37)	3.21* (0.52, 5.90)	4.07* (0.15, 7.99)
Group differences				
(Possesses ε4) - (no ε4)	–0.19 (–3.80, 3.42)	–0.28 (–3.40, 2.84)	–0.73 (–5.12, 3.66)	
Pseudo-R^2	0.90	0.80	0.82	0.86

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

Table 3: (*Continued*)

Parameter	General cognitive performance	Memory	Executive functioning	Language
	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Baseline cognitive status				
Quartile 1 (n = 1036)	2.75* (1.75, 3.75)	2.54* (1.44, 3.64)	1.17* (0.21, 2.13)	1.85* (0.65, 3.05)
Quartile 2 (n = 1013)	0.42 (–0.50, 1.34)	0.37 (–1.14, 1.88)	0.47 (–0.75, 1.69)	0.89 (–0.91, 2.69)
Quartile 3 (n = 1013)	–0.23 (–1.27, 0.81)	0.04 (–1.19, 1.27)	–0.17 (–1.99, 1.65)	–0.10 (–2.10, 1.90)
Quartile 4 (n = 1011)	0.58 (–1.85, 3.01)	0.46 (–3.36, 4.28)	0.45 (–1.65, 2.55)	1.16 (–2.43, 4.75)
Group differences				
Quartile1 - Quartile2	2.33* (0.96, 3.70)	2.17* (0.29, 4.05)	0.70 (–0.85, 2.25)	0.95 (–1.23, 3.13)
Quartile1 - Quartile3	2.98* (1.53, 4.43)	2.51* (0.86, 4.16)	1.35 (–0.71, 3.41)	1.95 (–0.38, 4.28)
Quartile1 - Quartile4	2.18 (–0.47, 4.83)	2.08 (–1.90, 6.06)	0.72 (–1.59, 3.03)	0.69 (–3.09, 4.47)
Quartile2 - Quartile3	0.65 (–0.74, 2.04)	0.33 (–1.61, 2.27)	0.65 (–1.53, 2.83)	0.99 (–1.70, 3.68)
Quartile2 - Quartile4	–0.16 (–2.77, 2.45)	–0.09 (–4.21, 4.03)	0.02 (–2.39, 2.43)	–0.26 (–4.28, 3.76)
Quartile3 - Quartile4	–0.80 (–3.45, 1.85)	–0.42 (–4.44, 3.60)	–0.62 (–3.38, 2.14)	–1.26 (–5.38, 2.86)
Pseudo-R^2	0.89	0.80	0.80	0.85
Cardiovascular risk burden				
Zero (n = 884)	4.30* (1.12, 7.48)	5.02* (1.86, 8.18)	2.46 (–0.62, 5.54)	4.33* (0.23, 8.43)
One (n = 1414)	4.97* (2.38, 7.56)	4.27* (1.47, 7.07)	4.24* (1.79, 6.69)	5.22* (2.14, 8.30)
Two (n = 1167)	4.15* (1.39, 6.91)	4.02* (0.84, 7.20)	3.14* (0.36, 5.92)	4.36* (0.99, 7.73)
Three or more (n = 608)	3.48 (–0.91, 7.87)	3.42 (–2.50, 9.34)	2.97 (–0.36, 6.30)	3.51 (–1.84, 8.86)
Group differences				
Zero - One	–0.68 (–4.78, 3.42)	0.75 (–3.46, 4.96)	–1.78 (–5.72, 2.16)	–0.88 (–6.02, 4.26)
Zero - Two	0.15 (–4.04, 4.34)	1.00 (–3.49, 5.49)	–0.68 (–4.82, 3.46)	–0.03 (–5.34, 5.28)
Zero - Three	0.82 (–4.59, 6.23)	1.60 (–5.10, 8.30)	–0.51 (–5.06, 4.04)	0.82 (–5.92, 7.56)
One - Two	0.83 (–2.95, 4.61)	0.25 (–3.98, 4.48)	1.10 (–2.60, 4.80)	0.86 (–3.71, 5.43)
One - Three	1.49 (–3.61, 6.59)	0.85 (–5.70, 7.40)	1.26 (–2.88, 5.40)	1.71 (–4.46, 7.88)
Two - Three	0.67 (–4.50, 5.84)	0.60 (–6.12, 7.32)	0.17 (–4.16, 4.50)	0.85 (–5.46, 7.16)
Pseudo-R^2	0.90	0.80	0.82	0.86

Note. Multilevel models of changes for general cognitive performance, memory, executive functioning, and language using time in study as the timescale. The retest parameters correspond to β_2 parameters in equation 4, and group differences correspond to parameter β_4. Retest effects are parameterized here as the jump in performance between the first and subsequent testing occasions.
*$p < 0.05$.

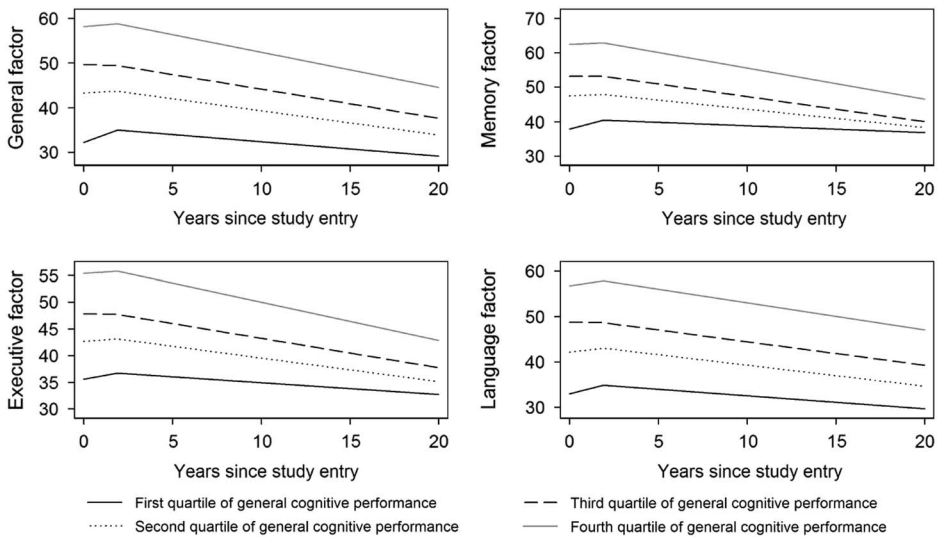

Fig. 1. Model-estimated trajectories of general and domain-specific cognitive functioning by baseline cognitive performance level: Results from WHICAP (N = 4073). Graphical depiction of model-estimated trajectories of general and domain-specific cognitive function to illustrate effect of practice effects in cognitive aging. The second study visit occurred on average 1.9 years after the first study visit, so the practice effect depicted is 1.9 years after study entry in figures.

https://doi.org/10.1017/S1355617715000508 Published online by Cambridge University Press

(Boyle, Wilson, Aggarwal, Tang, & Bennett, 2006). In this study, dementia diagnoses were in part based on a published algorithm using education-adjusted neuropsychological test scores (Stern et al., 1992). People whose neuropsychological test scores were consistent with dementia according to this algorithm, and whose daily function and level of independence deteriorated from previous levels according to self or informant report, were eligible for a diagnosis of dementia. In WHICAP and other similar epidemiologic cohorts, consensus diagnoses of dementia have been found to be meaningfully associated with declining cognitive trajectories and biomarkers for AD and cerebrovascular disease. The consensus group is blind to previous diagnosis in WHICAP, so although the criteria for dementia remain stable at each visit, there is no way to ensure continuity of diagnosis if people who had low scores at their initial visit rise slightly above the cut score on one or a few tests at follow-up.

This study indicates that retest effects do not differ by race/ethnicity or years of education, which were intended to be proxies for testing experience. However, years of education only captures testing experience from early life, and does not reflect experiences accumulated throughout life. Admittedly, race and ethnicity are imperfect markers of test experience, and thus our results cannot conclusively disprove the hypothesis that test experience plays a role in retest effects. Furthermore, most Hispanic participants in WHICAP were immigrants, whose years of education are systematically lower and do not easily translate to years of education in the United States (Hoffmeyer-Zlotnik et al., 2005).

The finding of differential retest effects by baseline cognitive status is likely attributable to regression to the mean (Barnett, van der Pols, & Dobson, 2005). Most participants performing in the lowest quartile of cognitive performance had a study diagnosis of dementia. Persons with dementia have impaired learning and memory, and thus one might expect they should exhibit smaller retest effects assuming that retest is attributable largely to episodic memory. Previous studies have reported no retest effects in persons with MCI and dementia (Darby et al., 2002; Schrijnemaekers et al., 2006) or minimal (Duff et al., 2011). Incipient dementia may not attenuate retest effects if procedural memory accounts for improvement on repeated test administration (Mitrushina, Boone, Razani, & D'Elia, 2005). Procedural memory, or long-term, unconscious recollection of previous experiences important for retaining skills (e.g., typing on a keyboard or riding a bicycle), is relatively well-preserved in people with dementia (Meyer & Schvaneveldt, 1971; Perani et al., 1993; Sabe, Jason, Juejati, Leiguarda, & Starkstein, 1995; Schaie, 2005; Tulving & Markowitsch, 1998). This rationale may be limited to measures in which procedural memory has greater influence; tests of confrontation naming and verbal comprehension in the language factor are less susceptible to this reasoning. Indeed, we only found retest differences by baseline cognitive performance for the general cognitive performance and memory factors, not the executive functioning or language factors (Table 3).

Limitations of our study must be noted. First, we defined retest effects in two ways based on the discontinuity between first and second assessments, and on the square root of the number of prior test occasions. The former approach imposes the assumption that the retest benefit is constant across the second and subsequent assessments. The latter approach assumes accumulating retest effects at each successive assessment, with diminishing additional benefit at each successive assessment. Although modest violations of either of these assumptions are plausible, such violations are unlikely to substantively alter our findings. There are other plausible specifications of retest effects. For example, if each successive test occasion were to hypothetically confer a slightly larger retest benefit, our effect estimates would be a weighted average of these effects. This phenomenon would obscure subgroup differences in the magnitude of retest effects only if such differences occurred for some, but not all, waves of assessment. We think such a complex pattern of retest effects is unlikely.

A second limitation is, regardless of how we parameterize them, retest effects are difficult to disentangle from aging in studies that have roughly equally spaced assessment intervals because the number of prior assessments is nearly collinear with time since baseline (Hoffman, Hofer, & Sliwinski, 2011). This challenge is common to most longitudinal studies of cognition. In the absence of random assignment of timing of the first assessment, simplifying assumptions are necessary to identify retest effects in studies with test–retest intervals longer than approximately a week (Hoffman et al., 2011; Sliwinski et al., 2010). Studies with very short test–retest intervals are optimal for distinguishing retest effects from normal cognitive aging because one can infer real change has not happened between the intervals (Salthouse & Tucker-Drob, 2008). We did not attempt to estimate retest differences as a function of the amount of time elapsed between successive tests because such variability is relatively small in WHICAP, by design, and any variance that is observed may be due to other variables such as respondents' health status or enthusiasm for participating in cognitive assessments. Because of this structural limitation in the test–retest intervals in WHICAP, our estimated retest effects are likely conservative because some declines due to aging are expected.

A third limitation is that, in our study, we cannot know for certain whether we are capturing retest effects between the first and second visits, or change in cognitive performance. Improvement in cognitive performance is unlikely given that many who showed larger retest effects had dementia, and cognition is not expected to improve over time in people with dementia. The retest effects in our regressions are based either on a contrast between cognitive performance at the first assessment and cognitive performance at subsequent assessments, or an accumulating benefit with diminishing returns. Thus, a further limitation of our approach is that, to the extent age-related change is incorrectly estimated, the estimated retest effect will also be incorrect (Hoffman et al., 2011). However, in a typical cohort study design, we believe this approach is the best available strategy to estimate retest effects.

A final limitation is that the present analysis was restricted to cognitive domains tested in WHICAP. Measures of spatial ability, processing speed, and higher-level task-switching, for example, were not available. The mechanisms by which retest effects operate, and thus predictors of differential retest effects, may differ for different domains. A final study limitation is that our parameterization of retest effects implicitly assumes variance in the retest effect, but we did not formally incorporate random effects for retest. Ideally, with additional data it would be possible to describe the variance of the retest effect in a multilevel model incorporating random effects for the retest term.

Retest effects cannot be ignored in longitudinal research in cognitive aging because they may mask age-related cognitive decline and potentially distort tracking of disease progression and detection of decline (Ronnlund & Nilsson, 2006; Ronnlund, Nyberg, Backman, & Nilsson, 2005). The present study empirically evaluated differential patterns of retest effects for several cognitive domains in a diverse sample of community-living older adults. Because we found no differential retest effects among observable demographic groups, our findings suggest that, although retest effects must be taken into account, differential retest effects may not limit the generalizability of inferences across groups in longitudinal research. Our study provides evidence that a commonly recognized bias may not be all that worrisome empirically. Although the findings suggest differential retest effects may not limit the generalizability of inferences across groups in longitudinal research, replication in other cohorts with different participant characteristics and retest intervals is warranted.

ACKNOWLEDGMENTS

The WHICAP study is supported by a National Institutes of Health grant (R01 AG037212 to Mayeux). This work was supported by UL1TR000040. We gratefully acknowledge a conference grant from the National Institute on Aging (R13AG030995 to D. Mungas) that facilitated data analysis for this project. Dr. Gross was supported by National Institutes of Health grant R03 AG045494-01. Dr. Benitez was supported by the Litwin Foundation and National Institutes of Health grants (UL1 TR000062 and KL2 TR000060). Dr. Shih was supported by an NIA R01 grant (R01AG043960). Dr. Bangen was supported by a National Institutes of Health grant (T32 MH1993417). Dr. Skinner was supported by an Alzheimer Disease Training Grant (5T32AG000258). Dr. Manly was supported by a National Institutes of Health grant (R01 AG028786 to Manly). No authors claim any conflicts of interest.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1355617715000508.

REFERENCES

Abner, E.L., Dennis, B.C., Mathews, M.J., Mendiondo, M.S., Caban-Holt, A., & Kryscio, R.J., … Investigators. (2012). Practice effects in a longitudinal, multi-center Alzheimer's disease prevention clinical trial. *Trials*, *13*, 217. doi: 10.1186/1745-6215-13-217.

Barnett, A.G., van der Pols, J.C., & Dobson, A.J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220.

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*, 118. doi: 10.1186/1471-2202-11-118

Basso, M.R., Bornstein, R.A., & Lang, J.M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist*, *13*(3), 283–292. doi: 10.1076/clin.13.3.283.1743

Baxter, L.C., Caselli, R.J., Johnson, S.C., Reiman, E., & Osborne, D. (2003). Apolipoprotein E epsilon 4 affects new learning in cognitively normal individuals at risk for Alzheimer's disease. *Neurobiology of Aging*, *24*(7), 947–952. doi: 10.1016/S0197-4580(03)00006-X

Ben-Yishay, Y., Diller, L., Mandleberg, I., Gordon, W., & Gerstman, L.J. (1974). Differences in matching persistence behavior during block design performance between older normal and brain-damaged persons: A process analysis. *Cortex*, *10*(2), 121–132. doi: 10.1016/S0010-9452(74)80003-1

Benedict, R.H., & Zgaljardic, D.J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 339–352. doi: 10.1076/jcen.20.3.339.822.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. doi: 10.1037/0033-2909.107.2.238

Blair, C.K., Folsom, A.R., Knopman, D.S., Bray, M.S., Mosley, T.H., & Boerwinkle, E., … Atherosclerosis Risk in Communities (ARIC) Study Investigators. (2005). APOE genotype and cognitive decline in a middle-aged cohort. *Neurology*, *64*(2), 268–276. doi: 10.1212/01.WNL.0000149643.91367.8A

Bontempo, D.E., & Hofer, S.M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A.D. Ong & M. van Dulmen (Eds.), *Handbook of methods in positive psychology* (pp. 153–175). New York: Oxford University Press.

Borenstein, A.R., Copenhaver, C.I., & Mortimer, J.A. (2006). Early-life risk factors for Alzheimer disease. *Alzheimer Disease and Associated Disorders*, *20*(1), 63–72. doi: 10.1097/01.wad.0000201854.62116.d7

Borsboom, D., Romeijn, J.W., & Wicherts, J.M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*, 75–98. doi: 10.1037/1082-989X.13.2.75

Boyle, P.A., Wilson, R.S., Aggarwal, N.T., Tang, Y., & Bennett, D.A. (2006). Mild cognitive impairment: Risk of Alzheimer disease and rate of cognitive decline. *Neurology*, *67*(3), 441–445.

Burke, E.F. (1997). A short note on the persistence of retest effects on aptitude scores. *Journal of Occupational and Organizational Psychology*, *70*, 295–301.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior*, *12*, 543–550. doi: 10.1016/S0022-5371(73)80034-9

Cagney, K.A., & Lauderdale, D.S. (2002). Education, wealth, and cognitive functioning in later life. *Journal of Gerontology: Psychological Sciences*, *2*, 163–172. doi: 10.1093/geronb/57.2.P163

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. doi: 10.1080/13854046.2012.680913

Collie, A., Maruff, P., Darby, D.G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of

neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society*, *9*(3), 419–428.

Cooper, D.B., Lacritz, L.H., Weiner, M.F., Rosenberg, R.N., & Cullum, C.M. (2004). Category fluency in mild cognitive impairment: Reduced effect of practice in test–retest conditions. *Alzheimer Disease and Associated Disorders*, *18*, 120–122.

Darby, D., Maruff, P., Collie, A., & McStephen, M. (2002). Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology*, *59*, 1042–1046. doi: 10.1212/WNL.59.7.1042

Del Ser, T., Gonzalez-Montalvo, J.-I., MartinezEspinosa, S., Delgado-Villapalos, C., & Bermejo, F. (1997). Estimation of premorbid intelligence in Spanish people with the Word Accentuation Test and its application to the diagnosis of dementia. *Brain and Cognition*, *33*, 343–356.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (2000). *California Verbal Learning Test*. (2nd ed.), San Antonio, TX: Psychological Corporation.

Dodge, H.H., Wang, C.N., Chang, C.C., & Ganguli, M. (2011). Terminal decline and practice effects in older adults without dementia: The MoVIES project. *Neurology*, *77*(8), 722–730. doi: 10.1212/WNL.0b013e31822b0068

Dodrill, C.B., & Troupin, A.S. (1975). Effects of repeated administrations of a comprehensive neuropsychological battery among chronic epileptics. *The Journal of Nervous and Mental Disease*, *161*(3), 185–190. doi: 10.1097/00005053-197509000-00006

Duff, K., Lyketsos, C.G., Beglinger, L.J., Chelune, G., Moser, D.J., Arndt, S., … McCaffrey, R.J. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *American Journal of Geriatric Psychiatry*, *19*(11), 932–939. doi: 10.1097/JGP.0b013e318209dd3a

Dugbartey, A.T., Townes, B.D., & Mahurin, R.K. (2000). Equivalence of the Color Trails Test and Trail Making Test in nonnative English-speakers. *Archives of Clinical Neuropsychology*, *15*(5), 425–431. doi: 10.1016/S0887-6177(99)00034-7

Ferrer, E., Salthouse, T.A., McArdle, J.J., Stewart, W.F., & Schwartz, B.S. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychology and Aging*, *20*(3), 412–422. doi: 10.1037/0882-7974.20.3.412

Ferrer, E., Salthouse, T.A., Stewart, W.F., & Schwartz, B.S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging*, *19*(2), 243–259. doi: 10.1037/0882-7974.19.2.243

Flicker, L. (2010). Cardiovascular risk factors, cerebrovascular disease burden, and healthy brain aging. *Clinics in Geriatric Medicine*, *26*(1), 17–27. doi: 10.1016/j.cger.2009.12.005

Frank, R., Wiederholt, W.C., Kritz-Silverstein, D.K., Salmon, D.P., & Barrett-Connor, E. (1996). Effects of sequential neuropsychological testing of an elderly community-based sample. *Neuroepidemiology*, *15*(5), 257–268. doi: 10.1159/000109915

Gibbons, L.E., Carle, A.C., Mackin, R.S., Harvey, D., Mukherjee, S., & Insel, P., … Alzheimer's Disease Neuroimaging Initiative. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior*, *6*(4), 517–527.

Glymour, M.M., & Manly, J.J. (2008). Lifecourse social conditions and racial and ethnic patterns of cognitive aging. *Neuropsychology Review*, *18*, 223–254. doi: 10.1007/s11065-008-9064-z

Glymour, M.M., Weuve, J., Berkman, L.F., Kawachi, I., & Robins, J.M. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemiology*, *162*(3), 267–278.

Gould, S.J. (1996). *The mismeasure of man*. New York: W.W. Norton & Company.

Gross, A.L., Jones, R.N., Fong, T.G., Tommet, D., & Inouye, S.K. (2014). Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology*, *42*, 144–153.

Gross, A.L., Sherva, R., Mukherjee, S., Newhouse, S., Kauwe, J.S.K., & Munsie, L.M., … AD Genetics Consortium. (2014). Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology*, *43*, 194–205.

Haan, M.N., Shemanski, L., Jagust, W.J., Manolio, T.A., & Kuller, L. (1999). The role of APOE epsilon4 in modulating effects of other risk factors for cognitive decline in elderly persons. *Journal of the American Medical Association*, *282*, 40–46. doi: 10.1001/jama.282.1.40

Harrison, J.E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *British Journal of Clinical Psychology*, *39*(Pt 2), 181–191. doi: 10.1348/014466500163202

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., & Gerrard, M.O.M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385.

Hayden, K.M., Jones, R.N., Zimmer, C., Plassman, B.L., Brown-dyke, 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*(2), 128–137. doi: 10.1097/WAD.0b013e3181ffa76d

Heilbronner, R.L., Sweet, J.J., Attix, D.K., Krull, K.R., Henry, G.K., & Hart, R.P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*, 1267–1278. doi:10.1080/13854046.2010.526785. doi: 10.1080/13854046.2010.526785

Hernan, M.A., & Robins, J.M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, *60*, 578–596. doi: 10.1136/jech.2004.029496

Hoffman, L., Hofer, S.M., & Sliwinski, M.J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychology and Aging*, *26*, 778–791. doi: 10.1037/a0023910

Hoffmeyer-Zlotnik, J.H.P., & Warner, U. (2005). How to measure education in cross-national comparison: Hoffmeyer-Zlotnik/Warner-Matrix of education as a new instrument. In J.H.P. Hoffmeyer-Zlotnik & J.A. Harkness (Eds.), *Methodological aspects in cross-national research* (pp. 223–240). ZUMA Nachrichten Special 11. Mannheim: ZUMA.

Horton, A.M. Jr. (1992). Neuropsychological practice effects x age: A brief note. *Perceptual and Motor Skills*, *75*(1), 257–258.

Howieson, D., Carlson, N., Moore, M., Wasserman, D., Abendroth, C., Payne-Murphy, J., & Kaye, J. (2008). Trajectory of mild cognitive impairment onset. *Journal of the International Neuropsychological Society*, *14*, 192–198.

Ivnik, R.J., Smith, G.E., Lucas, J.A., Petersen, R.C., Boeve, B.F., Kokmen, E., & Tangalos, E.G. (1999). Testing normal older

people three or four times at 1- to 2-year intervals: Defining normal variance. *Neuropsychology*, *13*, 121–127.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., & Dartigues, J.F. (1997). A five-year longitudinal study of Mini-Mental State Examination in normal aging. *American Journal of Epidemiology*, *145*, 498–506.

Jacobs, D.M., Sano, M., Dooneief, G., Marder, K., Bell, K.L., & Stern, Y. (1995). Neuropsychological detection and characterization of preclinical Alzheimer's disease. *Neurology*, *45*(5), 957–962. doi: 10.1212/WNL.45.5.957

Johnson, J.K., Gross, A.L., Pa, J., McLaren, D.G., Park, L.Q., & Manly, J.J., for the Alzheimer's Disease Neuroimaging Initiative (2012). Longitudinal change in neuropsychological performance using latent growth models: A study of mild cognitive impairment. *Brain Imaging and Behavior*, *6*(4), 540–550. doi: 10.1007/s11682-012-9161-8

Jones, R.N., Rudolph, J.L., Inouye, S.K., Yang, F.M., Fong, T.G., Milberg, W.P., … Marcantonio, E.R. (2010). Development of a unidimensional composite measure of neuropsychological functioning in older cardiac surgery patients with good measurement precision. *Journal of Clinical and Experimental Neuropsychology*, *32*, 1041–1049.

Juster, F.T., & Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, *30*(Suppl.), 7–56. doi: 10.2307/146277

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

Kausler, D.H. (1994). *Learning and memory in normal aging*. San Diego, CA: Academic Press.

Laird, N.M., & Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974. doi: 10.2307/2529876.

Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., … Willis, R.J. (2005). The Aging, Demographics, and Memory Study: Study design and methods. *Neuroepidemiology*, *25*, 181–191. doi: 10.1159/000087448

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Luchsinger, J.A., Tang, M.-X., Shea, S., & Mayeux, R. (2001). Diabetes mellitus and risk of Alzheimer's disease and dementia with stroke in a multiethnic cohort. *American Journal of Epidemiology*, *154*, 635–641. doi: 10.1093/aje/154.7.635

Luchsinger, J.A., Reitz, C., Honig, L.S., Tang, M.-X., Shea, S., & Mayeux, R. (2005). Aggregation of vascular risk factors and risk of incident Alzheimer's disease. *Neurology*, *65*, 545–551. doi: 10.1212/01.wnl.0000172914.08967.dc

Machulda, M.M., Pankratz, V.S., Christianson, T.J., Ivnik, R.J., Mielke, M.M., Roberts, R.O., … Petersen, R.C. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *The Clinical Neuropsychologist*, *27*(8), 1247–1264. doi: 10.1080/13854046.2013.836567

Manly, J.J., Byrd, D.A., Touradji, P., & Stern, Y. (2004). Acculturation, reading level, and neuropsychological test performance among African American elders. *Applied Neuropsychology*, *11*(1), 37–46.

Manly, J.J., Jacobs, D.M., Sano, M., Bell, K., Merchant, C.A., ... Stern, Y. (1999). Effect of literacy on neuropsychological test performance in nondemented, education-matched elders. *Journal of the International Neuropsychological Society*, *5*(3), 191–202.

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*(3), 341–348.

Manly, J.J., Schupf, N., Tang, M.-X., & Stern, Y. (2005). Cognitive decline and literacy among ethnically diverse elders. *Journal of Geriatric Psychiatry and Neurology*, *18*, 213–217. doi: 10.1177/0891988705281868

Manly, J.J., Tang, M.X., Schupf, N., Stern, Y., Vonsattel, J.P., & Mayeux, R. (2008). Frequency and course of mild cognitive impairment in a multiethnic community. *Annals of Neurology*, *63*(4), 494–506.

Manly, J.J., Touradji, P., Tang, M.X., & Stern, Y. (2003). Literacy and memory decline among ethnically diverse elders. *Journal of Clinical and Experimental Neuropsychology*, *25*(5), 680–690.

Mann, V.A., Sasanuma, S., Sakuma, N., & Masaki, S. (1990). Sex differences in cognitive abilities: A cross-cultural perspective. *Neuropsychologia*, *28*(10), 1063–1077. doi: 10.1016/0028-3932(90)90141-A

Mattis, S. (1988). *Dementia Rating Scale: Professional manual*. Odessa, FL: Psychological Assessment Resources.

McCaffrey, R.J., Onega, A., Orsillo, S.M., Nelles, W.B., & Haase, R.F. (1992). Practice effects in repeated neuropsychological assessments. *The Clinical Neuropsychologist*, *6*, 32–42. doi: 10.1080/13854049208404115

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. doi: 10.1037/h0031564

Mitrushina, M., Boone, K.B., Razani, J., & D'Elia, L.F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology*, *47*(6), 790–801. doi: 10.1002/1097-4679(199111)47:6 < 790::AID-JCLP2270470610 > 3.0.CO;2-C

Mungas, D., Reed, B.R., Marshall, S.C., & González, H.M. (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology*, *14*(2), 209–223.

Muthén, B.O., & Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371–402. doi: 10.1037//1082-989X.2.4.371

Muthén, L.K., & Muthén, B.O. (1998–2012). *Mplus User's Guide* (7th ed.), Los Angeles, CA: Muthén & Muthén.

Perani, D., Bressi, S., Cappa, S.F., Vallar, G., Alberini, M., Grassi, F., … Fazio, F. (1993). Evidence of multiple memory systems in the human brain: A [18F] FDG PET metabolic study. *Brain*, *116*, 903–919. doi: 10.1093/brain/116.4.903

Rabbitt, P.M. (1993). Does it all go together when it goes? *Quarterly Journal of Experimental Psychology*, *46*(A), 385–433. doi: 10.1080/14640749308401055

Rabbitt, P., Diggle, P., Smith, D., Holland, F., & McInnes, L. (2001). Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologia*, *39*(5), 532–543. doi: 10.1016/S0028-3932(00)00099-3

Rabbitt, P., Diggle, P., Holland, F., & McInnes, L. (2004). Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *59*(2), 84–97. doi: 10.1093/geronb/59.2.P84

Rabbitt, P., Lunn, M., Wong, D., & Cobain, M. (2008). Age and ability affect practice gains in longitudinal studies of cognitive change. *The Journals of Gerontology. Series B,*

Psychological Sciences and Social Sciences, *63*(4), P235–P240. doi: 10.1093/geronb/63.4.P235

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.), Thousand Oaks, CA: Sage.

Reitan, R. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*, 271–276. doi: 10.2466/PMS.8.7.271-276

Ronnlund, M., & Nilsson, L.G. (2006). Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence*, *34*, 63–78. doi:10.1016/j.intell.2005.06.004

Ronnlund, M., Nyberg, L., Backman, L., & Nilsson, L. (2005). Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study. *Psychology and Aging*, *20*, 3–18. doi: 10.1037/0882-7974.20.1.3

Sabe, L., Jason, L., Juejati, M., Leiguarda, R., & Starkstein, S.E. (1995). Dissociation between declarative and procedural learning in dementia and depression. *Journal of Clinical and Experimental Neuropsychology*, *17*, 841–848. doi: 10.1080/01688639508402433

Salthouse, T.A. (2009). When does age-related cognitive decline begin? *Neurobiology and Aging*, *30*(4), 507–514. doi: 10.1016/j.neurobiolaging.2008.09.023

Salthouse, T.A. (2010a). *Major issues in cognitive aging*. New York: Oxford University Press.

Salthouse, T.A. (2010b). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*, *24*, 563–572. doi: 10.1037/a0019026

Salthouse, T., Schroeder, D., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology*, *40*(5), 813–822. doi: 10.1037/0012-1649.40.5.813

Salthouse, T.A., & Tucker-Drob, E.M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, *22*, 800–811. doi: 10.1037/a0013091

Schaie, K.W. (2005). *Developmental influences on adult intelligence: The Seattle Longitudinal Study*. New York: Oxford University Press.

Schneider, B.C., Gross, A.L., Bangen, K.J., Skinner, J.C., Benitez, A., Glymour, M.M., … Luchsinger, J.A. (2014). Association of vascular risk factors with cognition in a multiethnic sample. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *70*(4), 532–544.

Schofield, P.W., Mosesson, R., Stern, Y., & Mayeux, R. (1995). The age at onset of Alzheimer's disease and an intracranial area measurement: A relationship. *Arch Neurol*, *52*(1), 95–98.

Schrijnemaekers, A.M., de Jager, C.A., Hogervorst, E., & Budge, M.M. (2006). Cases with mild cognitive impairment and Alzheimer's disease fail to benefit from repeated exposure to episodic memory tests as compared with controls. *Journal of Clinical and Experimental Neuropsychology*, *28*, 438–455. doi: 10.1080/13803390590935462

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*(3), 402–411. doi: 10.1037/a0017515

Singer, J.D., & Willet, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Sliwinski, M., Hoffman, L., & Hofer, S. (2010). Modeling retest and aging effects in a measurement burst design. In P.C.M. Molenaar & K.M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (pp 37–50). Washington, DC: American Psychological Association.

Spreen, O., & Benton, A. (1969). *Neurosensory Centre Comprehensive Examination for Aphasia*. Victoria, British Columbia, Canada: University of Victoria.

Steiger, J.H., & Lind, J.C. (1980). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society; Iowa City, IA.

Stern, Y., Andrews, H., Pittman, J., Sano, M., Tatemichi, T., Lantigua, R., & Mayeux, R. (1992). Diagnosis of dementia in a heterogeneous population. Development of a neuropsychological paradigm-based diagnosis of dementia and quantified correction for the effects of education. *Archives of Neurology*, *49*, 453–460.

Stern, Y., Gurland, B., Tatemichi, T.K., Tang, M.X., Wilder, D., & Mayeux, R. (1994). Influence of education and occupation on the incidence of Alzheimer's disease. *Journal of the American Medical Association*, *271*, 1004–1010. doi: jama.271.13.1004

Stuss, D., Stethem, L., & Poirier, C. (1987). Comparison of three tests of attention and rapid information processing across six age groups. *The Clinical Neuropsychologist*, *1*, 139–152.

Tang, M.-X., Cross, P., Andrews, H., Jacobs, D.M., Small, S., Bell, K., … Mayeux, R. (2001). Incidence of AD in African-Americans, Caribbean Hispanics, and Caucasians in northern Manhattan. *Neurology*, *56*, 49–56. doi: 10.1212/WNL.56.1.49

Tang, M.X., Stern, Y., Marder, K., Bell, K., Gurland, B., Lantiqua, R., … Mayeux, R. (1998). The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *Journal of the American Medical Association*, *279*(10), 751–755. doi: 10.1001/jama.279.10.751

Thorndike, E.L. (1922). Practice effects in intelligence tests. *Journal of Experimental Psychology*, *5*, 101–107. doi: 10.1037/h0074568

Tulving, E., & Markowitsch, H.J. (1998). Episodic and declarative memory: Role of the hippocampus. *Hippocampus*, *8*(3), 198–204. doi: 10.1002/(SICI)1098-1063(1998)8:3 < 198::AID-HIPO2 > 3.3.CO;2-J

Van der Elst, W., Van Boxtel, M.P.J., Van Breukelen, G.J.P., & Jolles, J. (2008). Detecting the significance of changes in performance on the Stroop Color-Word Test, Rey's Verbal Learning Test, and the Letter Digit Substitution Test: The regression-based change approach. *Journal of the International Neuropsychological Society*, *14*, 71–80. doi: 10.10170S1355617708080028

Voyer, D., Voyer, S., & Bryden, M.P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2), 250–270. doi: 10.1037//0033-2909.117.2.250

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. New York: The Psychological Corporation.

Wilkinson, G.S., & Robertson, G.J. (2006). *Wide Range Achievement Test 4 professional manual*. Lutz, FL: Psychological Assessment Resources.

Wilson, R.S., Li, Y., Bienias, J.L., & Bennett, D.A. (2006). Cognitive decline in old age: Separating retest effects from the effects of growing older. *Psychology and Aging*, *21*(4), 774–789. doi: 10.1037/0882-7974.21.4.774

Wilson, R., Leurgans, S., Boyle, P., & Bennett, B. (2011). Cognitive decline in prodromal Alzheimer disease and mild cognitive impairment. *Archives of Neurology*, *68*, 351–356.

Zehnder, A., Blasi, S., Berres, M., Spiegel, R., & Monsch, A. (2007). Lack of practice effects on neuropsychological tests as early cognitive markers of Alzheimer disease? *American Journal of Alzheimer's Disease and Other Dementias*, *22*, 416–426.