Measurement Equivalence of the Neuropsychological Test Battery of the Canadian Study of Health and Aging across Two Levels of Educational Attainment

by

Paul W. H. Brewster

B. A., York University (2009)

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

© Paul Brewster, 2011
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy
or other means, without the permission of the author.

Supervisory Committee

Measurement Equivalence of the Neuropsychological Test Battery of the Canadian Study of Health and Aging across Two Levels of Educational Attainment

by

Paul W. H. Brewster

B. A., York University (2009)

Supervisory Committee

Dr. Stuart W. S. MacDonald, (Department of Psychology)
Supervisor

Dr. Holly A. Tuokko, (Department of Psychology)
Departmental Member

Abstract

Supervisory Committee
Dr. Stuart W. S. MacDonald, (Department of Psychology)
Supervisor
Dr. Holly A. Tuokko, (Department of Psychology)
Departmental Member

Objective. This thesis examines the invariance of a battery of neuropsychological tests to known education-associated differences in strategy implementation and neural resource allocation underlying cognitive task performance in older adults without cognitive impairment.

Methods. Confirmatory factor analysis was used to evaluate the fit of a three-factor measurement model (Verbal Ability, Visuospatial Ability, Long-term Retention; Tuokko et al., 2009) to scores from the neuropsychological battery of the Canadian Study of Health and Aging (CSHA) for the purpose of confirming the latent constructs measured by its 11 tests. Measurement equivalence of the model across lower- (LE; ≤8 years) and higher-educated (HE; ≥9 years) participants was then evaluated using invariance testing.

Results. The measurement model demonstrated adequate fit across LE and HE samples but the loadings of the 11 tests (indicators) onto the three factors could not be constrained equal across groups. Two non-invariant tests of verbal ability (Animal fluency, token test) were identified that, when freed from constraints, produced an invariant model. Constraint of factor covariances did not compromise the partial invariance of this model. Because demographic characteristics of the LE and HE samples differed significantly, findings were replicated on age-and sex-matched subsamples.

Conclusions. Two measures of verbal ability were not invariant across HE and LE samples of older adults, suggesting that the cognitive processes underlying performance on these tests may vary as a function of educational attainment.

Table of Contents

Supervisory Committee ii
Abstract iii
Table of Contents iv
List of Tables vi
List of Figures vii
Acknowledgments viii
Introduction 1
Education and Neuropsychological Test Performance 2
Education and Strategy Implementation During Cognitive Performance 3
Education-Associated Differences in Structural and Functional Underpinnings of Cognitive Performance 5
Invariance of Neuropsychological Tests to Educational Attainment 12
Implications for Clinical and Experimental Neuropsychology 18
Objectives and Hypotheses 19
Methods 21
Participants 22
Measures 24
Neuropsychological Tests 24
Demographic Variables 29
Measurement Model 30
Statistical Analyses 31
Data Preparation 31
Confirmatory Factor Analysis 32
Invariance Testing 39
Results 44
Participant Characteristics 44
Data Preparation 48
Fit of the Measurement Model to CSHA-2 Neuropsychological Test Scores 51
Metric Invariance of the Measurement Model across HE and LE Samples 54
Discussion 57
Model Selection 59
Model Fit: Configural Invariance 62
Metric Invariance 64
Animal Fluency as a Non-Invariant Measure of Verbal Ability 67
The Token Test as a Non-Invariant Measure of Verbal Ability 69
Between-Group Age and Sex Differences 70
Potential Limitations 72
Conclusions 74
References 76
Appendix A 90
Appendix B 93
Appendix C 97

Appendix D 99
Appendix E 102

List of Tables

Table 1) Fit of the Measurement Model to CSHA-1 Data 31
Table 2) Characteristics of the study sample vs. the full CSHA-2 cohort 44
Table 3) Demographic characteristics and mean test scores of the study sample 47
Table 4a) Characteristics of the age-matched samples 49
Table 4b) Characteristics of the sex-matched samples 50
Table 5) Univariate and multivariate characteristics of the 11 observed variables in the full study sample 50
Table 6) Configural invariance: Baseline model fit in the full LE and HE samples 51
Table 7) Squared multiple correlations, standardized and unstandardized coefficients for the model when applied to the full HE and LE samples. 53
Table 8) Fit of the baseline and nested models to the study sample 54
Table 9) Metric invariance of the measurement model across the full LE and HE samples 55

List of Figures

Figure 1) Measurement Model of the CSHA Neuropsychological Test Battery. 30
Figure 2) Distribution of years of formal schooling in the HE sample..... 46
Figure 3) Distribution of years of formal schooling in the LE sample. 46

Acknowledgments

The CSHA was coordinated through the University of Ottawa and the Federal Government`s Laboratory Centre for Disease Control. The contents of this thesis proposal have not been published elsewhere. The results of a preliminary analysis were accepted for presentation at the 2011 meeting of the International Neuropsychological Society.

Data reported in this article were collected as part of the second wave of the CSHA. Waves 1 and 2 of the CSHA core study were funded by the Seniors' Independence Research Program, through Health Canada's National Health Research and Development Program (NHRDP project # 6606-3954-MC (S)). Additional funding was provided by Pfizer Canada Incorporated, through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP (project # 6603-1417-302R)), by Bayer Incorporated, and by the British Columbia Health Research Foundation (projects # 38 (93-2) and #34 (96-1)). The CSHA was coordinated through the University of Ottawa and the Federal Government`s Laboratory Centre for Disease Control. PWHB was supported in part by a Canada Graduate Scholarship Master`s Award from the Canadian Institutes for Health Research (2009-2010).

Introduction

With growing awareness of the social and economic implications of global population aging has come increased recognition of the importance of cognitive impairment and dementia as predictors of functional dependence, institutionalization and mortality in older adults. Of particular interest to health researchers are factors associated with reduced risk of pathological outcomes in old age, one of which is educational attainment early in life. Education contributes to the ascertainment and maintenance of cognitive abilities across the adult lifespan and buffers against cognitive decline and clinical expression of neuropathology in old age.

The protective effects of education are intriguing to health researchers because they suggest that intellectual engagement can provide the brain with a mechanism to combat expression of cognitive impairment. Both passive (ascertainment bias (Tuokko et al., 2003)) and active (cognitive reserve (Stern, 2009)) explanations have been proposed as a framework for understanding the protective effects of education on cognitive performance, and this remains an active area of research in cognitive neuroscience. Much of this interest stems from the premise that furthering understanding of the adaptive processes recruited by the damaged or diseased brain could lead to innovations in neurorehabilitation for individuals suffering from neurological disorders.

Epidemiologic research has contributed to our understanding of the association between education and cognitive impairment and dementia and continues to provide opportunities for testing theories of ascertainment bias and cognitive reserve. Neuropsychological tests, when available, serve an important role in these investigations because they provide a more sophisticated behavioural measure of brain function than

can be obtained by global cognitive screening measures such as the Mini Mental Status Examination (MMSE). Many neuropsychological tests have the added advantage of a research base that provides evidence of the psychological constructs that the tests measure and their neural correlates.

Although the positive association between educational attainment and neuropsychological tests performance has long been documented, less attention has been directed toward the potential moderating effects of education on the brain regions and psychological constructs that are recruited during test performance. This is an important question because many studies of ascertainment bias and cognitive reserve rely on the premise that the neuropsychological test scores that are included in their analyses provide an equivalent measure of brain function across groups.

Education and Neuropsychological Test Performance

The normative datasets that are used clinically to interpret neuropsychological test performance reveal the strength of the association between education and cognitive performance. Described as "potent and pervasive" by Lezak (2004), education is associated with significantly higher scores on "global" cognitive ability indices such as the General Neuropsychological scale of the Halstead-Reitan battery (Karzmark et al., 1984) and Full Scale IQ on the Wechsler Adult Intelligence Scales. The tests included in the Mayo Older American Normative Study (WRAT-3 Reading, Boston Naming Test, Controlled Oral Word Association Test, Category Fluency, Rey-Osterrieth Complex Figure, Visual Form Discrimination, Trail Making Test A & B) all revealed strong associations with educational attainment (McCarthy et al., 2003; Ivnik et al., 1996) . Even those tests which ostensibly measure lower-level cognitive processes (e.g., line

cancellation and visual reproduction) have been shown to yield robust associations with education (Brucki & Nitrini, 2008; Le Carret et al., 2003). These investigations point toward a strong and clinically significant positive association between educational attainment and performance on neuropsychological tests.

Education and Strategy Implementation during Cognitive Performance

There is psychometric evidence that education-associated differences in scores on neuropsychological tests are a reflection of differential strategy implementation and cognitive processes underlying test performance. For example, Le Carret and colleagues (2003) examined whether the impact of educational attainment on nonverbal recognition memory was an artefact of differences in visuospatial ability. Analyses were conducted in a sample of cognitively intact older adults (mean age = 77) who completed the recognition subtest of the Benton Visual Retention Test and a visual discrimination task (Le Carret et al., 2003b). By first examining the association between level of education (no schooling/primary school/high school/post-secondary) and performance on the Benton Visual Retention Test, and subsequently adjusting their analysis by including performance on the visual discrimination task in the model as a covariate, the authors attempted to delineate the retrieval and visual search demands of the task. Their results indicated that the associations between education and performance on tests of nonverbal recognition memory are mediated by superior visuospatial perceptual abilities in individuals with higher levels of education, but that the association between education and nonverbal recognition memory retained significance and is thus not an artefact of education-associated differences in visuospatial ability. Supporting these findings of an effect of education on visual discrimination, an investigation of the effects of literacy on

visual search strategies in performance on a cancellation task found that literate individuals had more organized search strategies than those who were illiterate (Brucki & Nitrini, 2008). Illiterate individuals searched for targets randomly without recruiting the occidental (horizontal and left-to-right) scanning search strategies that were used by those who were literate. The more efficient search strategies of the literate individuals corresponded with significantly higher performance on the cancellation task compared to the illiterate group.

Le Carret and colleagues (Le Carret et al., 2003a) also examined education-associated differences in cognitive performance by applying a principal component analysis of a battery of neuropsychological tests administered to a sample of cognitively intact older adults (mean age = 77) and subsequently examining the association of each component with educational attainment (no schooling/primary school/high school/post-secondary). Of five components, two were significantly associated with educational attainment: “conceptualization ability” and “controlled processes”. The investigators interpreted their findings as evidence that the impact of education on cognitive performance is driven by the superior cognitive control and conceptualization ability of older adults with higher levels of educational attainment. Together, the results of the analyses conducted by Le Carret and colleagues (2003; 2003b) implicate higher-level frontal “executive” processes as responsible for the higher neuropsychological test scores of well-educated older adults relative to those with less schooling.

Education-Associated Differences in Structural and Functional Underpinnings of Cognitive Performance

More attention has been directed toward education-associated structural and functional differences in the brain. Structurally, education has been shown to mediate age-associated microstructural changes in the hippocampal formation (Piras et al., 2011). The diffusivity (motion) of water molecules in brain tissue provides a metric for quantifying structural integrity because microscopic structural barriers such as cell membranes, axons, and myelinsheaths impede the movement of water in brain tissue (Kantarci et al., 2001). Diffusion tensor imaging can thus be used to obtain an estimate of the microstructural integrity of brain structures based on their mean diffusivity. Piras and colleagues examined cross-sectionally whether the volumetry and mean diffusivity of bilateral deep grey matter structures, including the thalamus, caudate, globus pallidus, hippocampus and amygdala, were associated with educational attainment in a cognitively intact sample of 150 adults (mean age = 40, range: 18-65; mean years of education = 14, range: 5-21). The bilateral mean diffusivity of the hippocampus, but not the volumetry or mean diffusivity of other structures, was found to vary by educational attainment. This association retained its significance following adjustment for age, which was negatively associated with education and positively associated with diffusivity. Diffusivity reflects increased extracellular space due to cell death and synaptic loss (Kantarci et al., 2005), and the authors interpreted their finding of lower mean diffusivity in educated adults as evidence of "neural reserve" due to adult neurogenesis. These findings are consistent with investigations examining the effects of cognitive engagement on adult neurogenesis (Mirochnic et al., 2009) and synaptogenesis (Levi et al., 2003) using animal models.

The functional correlates of education-associated differences in cognitive performance have been investigated using functional MRI (fMRI) and event-related potentials (ERP). Springer, McIntosh, Winocur & Grady (2005) used fMRI to examine the relationship between educational attainment and neural activity in young and older populations during a test of recognition memory and how these differences affected task performance. As was expected, behavioural data revealed superior recognition memory in their young group (mean age = 23; mean years of education = 15) relative to the older group (mean age = 74; mean years of education = 14). In the young adults, educational attainment and performance accuracy were associated with the same pattern of activation during the recognition memory task. Higher education and higher recognition accuracy were associated with activation in the posterior cingulate gyrus, cuneus, precuneus, and lateral and medial temporal regions. Lower education and lower recognition accuracy were associated with activation in anterior regions including bilateral prefrontal cortex, premotor regions and the temporal pole of the left hemisphere. In contrast, there was a dissociation of patterns of activation associated with education and performance accuracy in the older adults. Higher education was correlated with increased right temporal and parietal cortex, cingulate gyrus and bilateral prefrontal activation, while lower education was associated with activation of the left postcentral and lingual gyri and the inferior and medial portions of the left temporal lobe. Higher recognition accuracy was associated with increased activation in the left premotor cortex, the right lingual gyrus and left caudate nucleus. Poorer recognition accuracy was related to activity in bilateral premotor regions and the middle temporal cortex. The authors interpreted their findings as evidence that increased frontal activity during recognition memory tasks is associated

with less education (and poorer performance) in young adults but higher education (and no association with performance) in older adults. The authors further suggested that their findings were consistent with the posterior-anterior shift theory of cognitive aging (Springer et al., 2005). This theory was built around evidence of increased frontal lobe activation in older adults vs. younger adults during cognitive performance. It is hypothesized that compensatory mechanisms that work to ameliorate age-associated cognitive decline are functions of the frontal lobes and that the age-associated shift in elevated functional activation from posterior brain regions to frontal regions during cognitive performance reflect the engagement of compensatory mechanisms by older adults (Davis et al., 2008). The findings of Springer and colleagues suggest that this shift is most pronounced in higher-educated older adults.

Other investigations have found an association between educational attainment, patterns of cortical activation, and cognitive performance in older adults. Angel and colleagues (2010) used ERP to examine the effects of cognitive aging on performance on a word-stem cued recall task, and further examined the effects of educational attainment on the association between age and task performance. Their behavioural findings indicated that education was associated with better cued recall in cognitively intact older adults (mean age = 66 years; high educated group mean = 14 years, low educated group mean = 9 years) but not in younger adults (mean age = 24 years; high educated group mean = 17 years, low educated group mean = 10 years). ERP data revealed cortical activation differences in younger adults such that those with more education demonstrated bilateral frontal and parietal lobe activation, whereas those with less education demonstrated bilateral frontal activation but only left parietal activation. The

higher-educated older adults demonstrated the same pattern of bilateral frontal and parietal activation as was seen in the educated younger adults. In contrast, the lower-educated older adults demonstrated parietal activation but no significant frontal activation. The authors interpreted these findings as evidence that education-associated differences in performance on tests of verbal memory in older adults are attributable to corresponding education-associated differences in recruitment of frontal lobe processes for item retrieval (Angel et al., 2010).

A complementary ERP investigation (Osorio et al., 2010) examined the neural correlates of word-stem completion under lexical and semantic priming conditions in highly educated younger (mean age = 26 years; mean years of education = 17) and older adults (mean age = 63; mean years of education = 17). Behavioural results of this investigation revealed superior word priming in the older vs. younger adults in both lexical and semantic priming conditions. ERP data revealed occipitoparietal lobe activation during task performance in both groups, with stronger activation in the young adult group vs. the older adult group. However, the older adult group additionally demonstrated significant bilateral frontal lobe activation. The authors interpreted their results as evidence of effective compensation for lower occipitoparietal lobe functioning in well-educated older adults via recruitment of additional frontal regions. The unexpected behavioural advantage of older adults relative to the younger adults was attributed to their corresponding baseline outperformance of the younger adults on the Mill Hill test, which measures verbal ability and crystallized knowledge and was included in this investigation as an index of IQ.

These studies provide evidence for a role of educational attainment in effectively

modifying the structure and function of neural networks underlying cognitive performance and corroborate the well-established behavioural findings of superior cognitive performance in higher vs. lower educated older adults. Many research groups have further investigated this association by characterizing groups based on their level of "cognitive reserve", which encompasses lifelong markers of cognitive engagement such as IQ and occupational complexity in addition to educational attainment.

Using a factor score combining educational attainment of participants and their performance on two measures of intelligence as a metric for cognitive reserve, Scarmeas and colleagues (2003) examined cognitive reserve-mediated brain activation during nonverbal recognition memory in younger (mean age = 23; mean years of education = 17) and older adults (mean age = 71; mean years of education = 15). Participants completed the recognition task while undergoing PET scanning and task difficulty was titrated to avoid group differences (old vs. young) in performance accuracy. Cognitive reserve was associated across age groups with the patterns of cortical activation that were elicited by the recognition task. However, the cognitive-reserve-associated patterns of activation differed as a function of age: relative to the younger group, cognitive reserve in the older group was associated with significantly less activation of the right inferior temporal and postcentral gyri, and significantly greater activation of the left cuneus. The authors interpreted their results as evidence that functional adaptation to age-related brain pathology in healthy older adults varies according to their level of cognitive reserve (Scarmeas et al., 2003).

Using the same dataset, the impact of cognitive reserve on topographic activation in response to increasing task demand was examined (Stern et al., 2005). A pattern of

activation was first identified in the younger group that corresponded with changes in task demand and was positively associated with cognitive reserve. The identified pattern included increased activation in the right lingual gyrus, inferior parietal lobe, association cortex, left posterior cingulate, and right and left calcarine cortex as well as decreased activation in right hippocampus, posterior insula, thalamus, and right and left operculum. This pattern of recruitment was the same in older adults with higher levels of cognitive reserve. However, older adults with lower levels of cognitive reserve demonstrated an opposite pattern of recruitment of this network (decreased activation in the right lingual gyrus, inferior parietal lobe, association cortex, left posterior cingulate, and right and left calcarine cortex as well as increased activation in right hippocampus, posterior insula, thalamus, and right and left operculum). The authors suggested that the network recruited by young adults and elders with high levels of cognitive reserve represented a neural manifestation of innate or acquired reserve.

Using fMRI, Bosch and colleagues (2010) examined the impact of cognitive reserve on activation of the default mode network in older adults (mean age = 73; education not reported) during a language comprehension task. The default mode network is a cluster of brain regions that are at the height of their activity during passive task conditions. When cognitive demands increase, a decrease in activation of the default mode network is observed that corresponds with increased activation of more task-specific brain regions (Buckner et al., 2008). Bosch and colleagues sought to determine whether cognitive reserve modulated activation of the default network, and whether individuals with higher levels of cognitive reserve would demonstrate higher or lower levels of activation of the default mode network during task performance relative to those

with a lower cognitive reserve. A factor score combining years of education, performance on a test of verbal IQ, self-reported leisure activities and occupational complexity was used as their proxy for cognitive reserve. Healthy older adults with higher levels of cognitive reserve showed less activation in tertiary brain regions responsible for language comprehension and a more consistent activation of the default mode network than those with lower cognitive reserve. The authors interpreted this finding as evidence that cognitive reserve in healthy older adults allows for more efficient usage of specialized brain networks by placing greater demands on default networks when engaging in cognitively demanding activities (Bosch et al., 2010).

Similarly, Bartres-Faz and colleagues (2009) investigated the interactions of cognitive reserve with regional brain anatomy and functional activation during performance of healthy older adults (mean age = 68; education not reported) on a test of working memory. A factor score combining years of education, performance on a test of verbal IQ, self-reported leisure activities and lifetime occupation was used as a proxy for cognitive reserve. Cognitive reserve was found to correlate negatively with functional activation in the inferior frontal lobes but this association was rendered nonsignificant after including the grey matter volume of this region (not independently associated with cognitive reserve) in the analysis as a covariate. The authors concluded that morphologic variations in functionally important regions may represent the "passive hardware" upon which cognitive reserve-associated "active software" relies (Bartres-Faz et al., 2009).

These findings collectively suggest that the neural correlates of cognitive processes vary as a function of an individual's level of education and other markers of

cognitive engagement, and that this association is independent of task difficulty. This association is further modulated by the structural integrity of associated brain regions.

Invariance of Neuropsychological Tests to Educational Attainment

The reviewed education-associated differences in strategy implementation and neural resource allocation that underlie differences in test performance suggest that neuropsychological tests may measure different cognitive processes and place demands on different neural networks as a function of the examinee's level of education. Although this issue has received very little attention in the literature, it is potentially problematic for behavioural studies that classify participants by educational attainment and subsequently compare performance accuracy across groups and draw inferences about the integrity of neural networks or cognitive abilities in low vs. higher educated individuals. In other words, educational attainment of examinees may compromise the measurement equivalence of neuropsychological tests.

Measurement equivalence (also termed measurement invariance) refers to the extent to which test items are perceived and interpreted similarly by test takers in different groups. More specifically, it refers to the demonstration that psychological constructs measured by a set of test scores are equivalent across all levels of a dependent variable (Byrne & Watkins, 2003). Establishing measurement equivalence is necessary in order to make meaningful group comparisons based on test scores because scores that do not provide an equivalent measure of the construct of interest across groups make it impossible to ascertain the neural and cognitive resources that were drawn upon by participants in each group during task performance (Vandenberg & Lance, 2000).

Measurement equivalence is most often investigated using factor analysis, although other methods (such as classical test theory and item response theory) are available (Teresi, 2006). Specifically, invariance testing of confirmatory factor analytic models is the method whereby factor analysis is used to evaluate measurement equivalence. Confirmatory factor analysis (CFA) allows for a hypothesis-testing approach to data analysis. It requires that investigators specify a measurement model (based on theory or prior research) by identifying latent constructs and the variables that are expected to provide a metric for these constructs. The fit of the specified model is then tested for validity in sample data. If the measurement model provides a strong fit to the sample data, then it is possible for investigators to interpret their model as a valid conceptualization of the latent factors measured by the variables included in their dataset. However, it is only accurate to conclude that the measurement model provides a valid conceptualization of their chosen dataset – the model may fit differently in an independent dataset if there are group differences in the demographic characteristics of participants. By identifying a measurement model, placing equality constraints on its parameters, and applying it to two or more samples of participants, it is possible to evaluate the measurement equivalence of the model across groups. By proxy, this approach provides a way of evaluating the measurement equivalence of the tests that were administered to participants.

When establishing the measurement equivalence of a chosen model there are a number of "levels" of invariance testing that can be tested. Specifically, models can be evaluated hierarchically based on their configural, metric, scalar and strict invariance. The level of invariance achieved by a model corresponds to the conclusions that can be

drawn regarding the measurement equivalence of the tests included in the model (Vandenberg & Lance, 2000). Configural invariance refers to the extent to which the patterns of free and fixed factor loadings imposed on the model indicators are equivalent across groups (Vandenberg & Lance, 2000). Configural invariance can be tested within the factor analytic framework by specifying the same number of factors and the same assignment of tests (indicators) to factors in each group. If the chosen measurement model represents a valid framework for the general psychological constructs that are measured by a set of tests (e.g., a three-factor model of intelligence), then the fit of this model when applied to data from a sample that recruits different psychological constructs for test performance (e.g., two factors underlying performance on the same test of intelligence) will result in poor model fit and prevent the establishment of configural invariance. If configural invariance cannot be demonstrated, then it makes no sense to conduct subsequent tests of measurement equivalence because completely different constructs are being recruited between groups and comparisons are meaningless. On the other hand, if configural invariance is demonstrated and the factor structure provides a comparable fit across two or more groups, this means that the same conceptual frame of reference can be used to understand test performance across groups and that comparisons can be drawn between groups regarding the loadings of model indicators on their respective factors. In this case it is appropriate to conduct further tests of measurement equivalence inasmuch as they are nested within the test of configural invariance (Brown, 2006).

If configural invariance is established, then the metric invariance of the model across groups can be evaluated. Whereas configural invariance tests the invariance of the

model factor structure, metric invariance evaluates the extent to which factor loadings can be constrained equal across groups (Vandenberg & Lance, 2000). Thus, metric invariance evaluates the strength of the association of each test (indicator) with the latent construct to which it is assigned. Examination of whether performance on a purported memory test places similar episodic memory demands across groups would represent an investigation of the metric invariance of this test according to the selected measurement model. Constraints on factor covariances and residual variances are also included in some definitions of metric invariance (Vandenberg & Lance, 2000).

If configural and metric invariance can be established then it is subsequently possible to investigate scalar invariance. Scalar invariance refers to the equivalence of item intercepts across groups (Vandenberg & Lance, 2000). In other words, scalar invariance evaluates differences in mean test scores across groups under the assumption that the test provides an equivalent measure of its latent factor. In cases where metric and scalar invariance are established, it is possible to conclude that the same psychological construct is measured on the same numerical scale in both groups. Scalar invariance is often investigated in cases where response bias or systematic group differences in performance are expected. However, examination of scalar invariance may be redundant in cases where mean-level group differences in test performance are known to exist.

In some cases model non-invariance is driven by many or all of the model indicators, but it can be caused by just one non-invariant indicator. When many or all indicators are non-invariant it is not possible to move forward with invariance testing or to make meaningful group comparisons regarding metric or scalar equivalence. However, if the minority of indicators are non-invariant, investigators have the option of testing the

partial invariance of their measurement model. Partial invariance testing involves systematically releasing model parameters and examining the ramifications of these adjustments on model fit (Brown, 2006). Factors associated with metric non-invariance can be identified by releasing all the factor loadings of a single factor, keeping the loadings of all other factors constrained equal, and examining the subsequent impact on model fit. If the model remains non-invariant after these modifications, then the metric invariance of that factor and its associated indicators can be established. Examination of each factor in the measurement model in this manner will reveal the non-invariant factors which, when released from constraint, result in an invariant measurement model. The specific indicators driving the non-invariance of a factor can then be isolated by sequentially applying constraints to each of the individual indicators that load on the non-invariant factor and examining the corresponding change in model fit. Indicators that cannot be constrained without rendering the measurement model non-invariant can be identified as the source of non-invariance. It is possible to subsequently continue with invariance testing by leaving the non-invariant indicators free to vary or by dropping them from the measurement model.

Different levels of invariance correspond to the types of comparisons that can be drawn across groups while remaining confident of measurement invariance (Vandenberg & Lance, 2000). Configural invariance is the weakest form of invariance and implies that groups have similar latent structures but could differ on other key dimensions. Metric invariance (also termed weak invariance) crudely establishes the construct validity of a set of tests across groups and provides evidence that tests provide a measure of the same psychological constructs in each group. Establishment of scalar invariance (also termed

strong invariance) allows investigators to assume that the same constructs are measured on the same numerical scale across groups. This is important because it allows for the valid comparison and interpretation of group mean differences in test performance. Finally, strict invariance is established when error variances are additionally held constant. Under these conditions, it is possible to conclude that observed group differences are due to differences at the latent variable level and not residual differences.

There has been only one investigation that has come close to examining the measurement equivalence of a cognitive test battery across levels of education. Paulo and Ryan (1994) employed exploratory factor analysis with varimax rotation to compare the factor structure of the Wechsler Adult Intelligence Scale (WAIS-R) in cognitively intact older adults age 75-95 with less than 12 years of formal education to its factor structure in a sample of older adults with 12 or more years of schooling. In older adults with less than 12 years of education a two-factor model (verbal ability, perceptual-organizational ability) emerged as the most parsimonious, while in those with 12 or more years of education a three-factor model (verbal comprehension, perceptual-organizational ability, freedom from distractibility) demonstrated the best fit. The authors concluded that the subtests of the WAIS-R measure different constructs in older adults with low vs. higher levels of education and suggested that clinicians be aware of these differences when interpreting test scores (Paolo & Ryan, 1994). The results of this analysis suggest that education so strongly modifies the psychological constructs measured by tests of cognitive ability that even configural invariance cannot be established.

Implications for Clinical and Experimental Neuropsychology

Issues surrounding the measurement equivalence of neuropsychological tests across levels of educational attainment are partly circumvented in clinical practice through the availability of education-stratified normative data for some tests. Stratified normative data allow clinicians to interpret test scores based on mean performance of those with similar levels of educational attainment. Thus, regardless of whether tests are measuring the same psychological constructs or placing demands on the same neural networks across different stratifications of educations, clinicians obtain an estimate of an examinee's performance relative to others with a similar educational background.

Unlike the clinical context, in research it is common (and often considered best practice) to compare group performance on neuropsychological tests using unadjusted raw scores. In cases where demographic-associated differences in cognitive performance are considered problematic, demographic variables are included as covariates in the statistical models along with the raw neuropsychological test scores of interest. This approach adjusts for mean differences in test scores between groups but does not account for potential group differences in the patterns of variance associated with test performance or the latent constructs that they measure. This is problematic because it is an assumption of many common statistical techniques, including analysis of variance (ANOVA) and regression, that independent variables are invariant across all levels of a dependent variable. It has even been argued that demonstration of measurement equivalence should precede all quantitative group comparisons (Baltes & Nesselroade, 1973). Although this requirement is rarely implemented, these statistical assumptions underscore the importance of establishing measurement equivalence of

neuropsychological tests prior to drawing education-associated inferences regarding group differences in brain structure or function.

Objectives and Hypotheses

Despite widespread interest in the association between education and cognitive functioning in older adulthood, there is no evidence that the neuropsychological tests that are used to evaluate cognitive ability are invariant across levels of education. In fact, evidence from exploratory factor analyses suggest the contrary, that education may compromise the psychological constructs measured by scores on a set of cognitive tests (Paolo & Ryan, 1994). The Canadian Study of Health and Aging (CSHA) presented the opportunity to test the hypothesis that neuropsychological test batteries measure the same latent constructs and cognitive processes in older adults regardless of their level of education. The factor structure of the CSHA neuropsychological test battery has been previously identified by Tuokko and colleagues (2009) in their investigation of the measurement equivalence of English and French neuropsychological test forms. This measurement model (described in greater detail in the Methods) demonstrated adequate model fit and its three-factor structure, measuring verbal ability, visuospatial ability and long-term retention, corresponds with other established models of cognitive ability (Flanagan & Harrison, 2005).

The purpose of this thesis is to examine whether the construct measurement of the CSHA neuropsychological test battery is invariant (or equivalent) across cognitively intact samples of older adults with low vs. higher levels of educational attainment. This purpose will be addressed by 1) fitting the measurement model identified by Tuokko and colleagues (2009) to samples of cognitively intact older adults with low vs. higher levels

of education, and 2) testing the configural and metric invariance of this measurement model to group differences in educational attainment. It is expected that the measurement model will demonstrate adequate fit in the study sample, thereby providing a workable representation of the psychological constructs measured by the CSHA neuropsychological battery. Based on prior research, full invariance of the measurement model is not an expected outcome. If confronted with non-invariance at the metric level, partial measurement invariance analyses will be pursued. To avoid redundancy, scalar invariance will only be investigated if mean scores on the tests included in the CSHA neuropsychological battery do not differ significantly between groups.

Methods

Study Design: The CSHA was a national longitudinal study of the epidemiology of dementia in Canada that included three waves of data collection: CSHA-1 (1991-1992), CSHA-2 (1995-1996) and CSHA-3 (2000-2001).

Representative samples of people aged 65 and over (N=10,263) were drawn from the 10 Canadian provinces in 1991 and 1992 (CSHA-1). Individuals over the age of 85 were oversampled in order to ensure that this age cohort was adequately represented in the CSHA. Following approval from institutional review boards and written informed consent from participants, the Modified Mini-Mental State Examination (Teng & Chui, 1987) was administered to identify candidates for clinical assessment. Those with 3MS scores lower than 78 plus all institutionalized residents received a clinical assessment, and a random sample of 494 participants with 3MS scores of 78 or higher also underwent clinical examination. A study physician conducted a physical and neurological examination of each participant and made a preliminary diagnosis. All participants who scored above 50 on the 3MS also completed a neuropsychological test battery administered by a psychometrist. Neuropsychological testing was only administered to participants with 3MS scores of ≥50 because those with scores lower than 50 were considered too cognitively impaired to undergo assessment. At a consensus case conference the study physician and neuropsychologist reviewed their preliminary diagnoses and a consensus diagnosis was established for each participant. In 1995-96, CSHA-2 used the same procedure to examine the incidence and progression of cognitive impairment and dementia in the cohort. At CSHA-3 (2001-02), all surviving participants not previously diagnosed with dementia were again screened using the 3MS and those

who scored ≤90 as well as all who had undergone a prior clinical assessment at CSHA-1 or CSHA-2 underwent diagnostic assessment following the same procedure as the previous waves of the CSHA. Full details of the CSHA can be found at www.csha.ca and elsewhere (McDowell et al., 2001).

Participants

Selection of Participants: All English-speaking participants who completed the neuropsychological test battery at CSHA-2 and were determined in the consensus case conference to have no cognitive impairment at that time were eligible for inclusion in the present analyses. Only English-speaking participants were included in this study because French and English neuropsychological test forms do not provide an equivalent measure of cognitive functioning (Tuokko et al., 2009). Inclusion was further limited to participants with no cognitive impairment because cognitive impairment and dementia have been shown to compromise the latent factor structure of other cognitive test batteries (Davis et al., 2003). In addition, cognitive impairment is more prevalent in populations with lower levels of education (McDowell et al., 2007) and therefore inclusion of cognitively impaired participants could obfuscate the findings of the present study if the distribution of cognitive impairment and dementia are not uniform across the low and higher-educated samples. Finally, otherwise eligible participants without data regarding years of formal education were necessarily excluded.

Classification of Participants: Participants were assigned to one of two groups based on their level of education: those with eight or fewer years of formal education were assigned to the low educated group (LE), and those with nine or more years of education were included in the higher educated group (HE). Low education was characterized as

eight or fewer years of education both to maximize the LE sample size and to adopt a meaningful cut point, as the completion of nine or more years of formal education typically corresponds with completion of one or more years of secondary school, while eight or fewer years of education indicates discontinuation of schooling prior to enrolment in secondary school. There is substantial inconsistency across studies regarding "cut points" for classifying LE vs. HE participants, but the cut point implemented here is consistent with some previous investigations (McDowell et al., 2007), (Hall et al., 1998).

Matched sampling: In order to attribute any potential study findings to the differences in educational attainment of the LE and HE samples it is necessary to ensure that groups are comparable on other demographic dimensions. In particular, significant baseline group differences in age or sex distributions would require matched sampling to ensure that findings are not an artefact of age or sex. This is important because both of these characteristics have been shown to independently compromise the measurement invariance of cognitive tests (Rodriguez-Aranda & Sundet, 2006). Adjustment for group differences in global cognitive status will not be implemented because all participants in the study sample underwent a physical examination by a physician and a cognitive assessment by a neuropsychologist and were determined to be cognitively intact. Matched sampling will be implemented instead of covarying for key demographic variables because it is an assumption of confirmatory factor analysis that a measurement model be invariant across all levels of a covariate, an assumption that cannot be met by sex or chronological age.

Measures

Neuropsychological Tests

The neuropsychological battery of the CSHA was designed to fulfill criteria for the detection of dementia according to the Diagnostic and Statistical Manual (DSM III-R), and has been described in detail elsewhere (Tuokko et al., 1995). The battery includes measures of abilities such as memory, language, abstract reasoning and visuospatial ability. Raw scores from the following tests were included in the present analyses:

Modified Buschke Cued Recall Test (BCRT): The BCRT is a measure of free and cued verbal recall (Tuokko & Crockett, 1989). This task involves first presenting examinees with a stimulus card depicting a random array of common objects. In response to semantic category prompts provided by the examiner (e.g., show me the vegetable), examinees are required to find and identify the appropriate object on the stimulus page. Subsequently, examinees are asked to recall each of the 12 pictures from memory. After the first free recall trial the examiner provides semantic cues for items that were missed in free recall, and the examinee is provided the opportunity to use the semantic cues to retrieve the missed items. This process is repeated three times (six trials total) and a delayed trial is administered after a five minute distracter task. Total free recall scores derived from the BCRT were examined in the present analyses. The BCRT has demonstrated sensitivity to memory deficits in older adults (Tuokko et al., 1991) and scores are moderately affected by education (Tuokko & Crockett, 1989).

Rey Auditory Verbal Learning Test (RAVLT): The RAVLT is a measure of verbal learning and memory (Bleecker et al., 1988). A classic list-learning test, it involves reading a list of 15 nouns to examinees and requesting immediate recall. Five sequential

trials of reading the same list to the examinee and requesting immediate recall are followed by a single administration of a new list of 15 nouns. After a 20-minute delay period examinees are again asked to recall from memory the word that were on the first list (administered in trials 1-5 but not re-administered after the delay). A recognition trial is administered immediately following the delayed recall trial which involves distinguishing the 15 words included in the first list from 35 distracter words. There are many scores that can be derived from performance on the RAVLT, but the measurement model identified by Tuokko and colleagues (2009) included the total free recall score, which consists of all five immediate recall trials of the first word list. The RAVLT is moderately associated with other measures of learning and memory, including the Logical Memory and Reproduction subtests of the Wechsler Memory Scale and the CVLT-II. It has demonstrated sensitivity to age-associated decline in memory acquisition and retention deficits in a vast array of neurological disorders (Dunlosky & Salthouse, 1996). The impact of education on RAVLT performance is disputed, with some investigations reporting a positive association (Van der Elst et al., 2005; Miatton et al., 2004), and others finding no association (Mitrushina et al., 1991). In the CSHA the delay period between immediate and delayed recall was abbreviated from 20 to 10 minutes.

Wechsler Memory Scale (WMS) Information subtest: The Information subtest of the WMS primarily measures orientation and crystallized knowledge by asking a series of questions regarding time (e.g., the date) place, personal information (e.g., date of birth) and current events (e.g., the current President of the United States; Wechsler & Stone, 1974). WMS Information has demonstrated sensitivity to prodromal dementia (Tierney et al., 2005) and performance is associated with education (Wechsler & Stone, 1974).

Wechsler Adult Intelligence Scale (WAIS-R) Block Design subtest: The Block Design subtest of the WAIS-R measures perceptual organization, visuospatial and motor skills (Wechsler, 1981). Examinees are provided with stimulus cards depicting two-colour geometric designs that they are instructed to replicate using coloured blocks. Scores are based on both accuracy and time to completion. Block Design scores is sensitive to age-related cognitive decline, brain injury and dementia, and correlates with the daily functioning and independent living of older adults (Lezak, 2004). Education does not contribute significantly to performance on this test (Kaufman et al., 2001). A Short Form of WAIS Block Design was used in the CSHA test protocol (Satz & Mogel, 1962).

Wechsler Adult Intelligence Scale (WAIS-R) Similarities subtest: The Similarities subtest of the WAIS-R provides a measure of verbal concept formation (Wechsler, 1981). Examinees are presented with word pairs (e.g., horse and frog) and asked to identify what the two words have in common. Scores are derived from the accuracy and quality of associations that examinees draw between the word pairs. The Similarities subtest is sensitive to age-related cognitive decline, brain injury (Warrington et al., 1986) and Alzheimer's disease (Fabrigoule et al., 1998). Education is strongly associated with Similarities subtest scores in older adults (McCarthy et al., 2003). A Short Form of WAIS Similarities was used in the CSHA test protocol (Satz & Mogel, 1962).

Wechsler Adult Intelligence Scale (WAIS-R) Comprehension subtest: The Comprehension subtest of the WAIS-R is a measure of verbal ability, judgement and remote memory (Lezak, 2004; Wechsler, 1981). Examinees are asked open-ended questions of common sense judgement, practical meanings and proverb meanings and scores are based on the accuracy and quality of responses. As a test that measures verbal

ability, the Comprehension subtest is sensitive to language deficits and left hemisphere lesions (Crosson et al., 1990), and has demonstrated sensitivity to deficits associated with Alzheimer disease and Multiple Sclerosis (Filley et al., 1989). Education is positively associated with performance on the Comprehension subtest (Heaton et al., 1996). A short form of WAIS Comprehension was used in the CSHA (Satz & Mogel, 1962).

Wechsler Adult Intelligence Scale (WAIS-R) Digit Symbol subtest: The Digit Symbol subtest of the WAIS-R measures psychomotor speed and sustained attention (Wechsler, 1981). The task involves copying nonsense symbols into squares. Each symbol is assigned a number and the task involves filling the correct symbol into squares with the matching number. Scores are based on the number of symbols accurately filled in within 90 seconds. Digit Symbol scores are sensitive to age-related cognitive decline (Ivnik et al., 1992) and many causes of brain dysfunction (Lezak, 2004). Education is associated with Digit Symbol scores in older adults (Mazaux et al., 1995; Lass et al., 1975).

Token Test: The Token Test (Benton & Hamsher, 1989) is a measure of receptive language and verbal comprehension. The task involves presenting the examinee with 20 plastic tokens in five colours, two shapes and two sizes (large vs. small) and asking the examinee to carry out 39 commands of increasing complexity with the tokens (e.g., "show me a triangle"..."Instead of the pink triangle, pick up the green square"). Scores on this test represent the number of commands that are correctly carried out by examinees, and are strongly associated with performance on other measures of comprehension and receptive language, including the Peabody Picture Vocabulary Test and Raven's Progressive Matrices (Lass et al., 1975). The Token Test has demonstrated sensitivity to receptive language difficulties in populations with traumatic brain injury and dementia

(Millis et al., 2001). Education-stratified normative data or evidence of the impact of education on performance on the Token test is not available.

Benton Visual Retention Test (BVRT): The BVRT (Benton, 1974) is a measure of visual memory, visual perception and visual-constructive abilities. The multiple-choice format of the BVRT used in the CSHA involves presenting examinees with stimulus cards that display geometric figures and subsequently requiring examinees to identify the stimulus geometric figure from a multiple choice card that includes the stimulus figure and three distracter figures. Scores on this test represent the percent of stimulus figures correctly recognized by participants over the 15 trials of the BVRT. The multiple choice administration of the BVRT loads primarily on factors representing memory and secondarily on factors representing attention and perceptual ability (Moses, Jr., 1986). The BVRT has demonstrated sensitivity to age-associated cognitive decline (Coman et al., 2002) and Alzheimer disease (Kawas et al., 2003). The association between BVRT scores and education falls within the moderate range, 0.40-0.60 (Le Carret et al., 2003a).

Phonemic Fluency: Phonemic fluency evaluates generative language production under restricted search conditions. In the CSHA, phonemic fluency was evaluated by having examinees generate as many words as possible within 60 seconds beginning with each of the letters F, A and S (Spreen & Benton, 1977). Scores reflect the number of words correctly generated over these three trials, and are associated with verbal IQ and processing speed (Boone et al., 1998), attentional control and working memory (Elias et al., 1997). Phonemic fluency has demonstrated sensitivity to age-associated cognitive decline (Sliwinski & Buschke, 1999), psychiatric disturbance (Henry & Crawford, 2005) and head injury (Henry & Crawford, 2004). Education is strongly associated with

performance on tests of phonemic fluency, and accounts for more variability in scores than age (Crossley et al., 1997). In a sample of older adults, those with the highest levels of educational attainment (13+ years) were found to generate twice the number of correct words as those with the lowest level of education (≤ 7 years).

Animal Fluency: Animal fluency evaluates the generation of language under semantically-controlled search conditions (Read, 1980). Animal fluency (referred to in other texts as category fluency or semantic fluency) was evaluated in the CSHA by having examinees name as many animals as they were able within 60 seconds. Like phonemic fluency, animal fluency scores correlate highly with performance on tests of attentional control and working memory (Rosen & Engle, 1997). Although associations with semantic memory have been documented for both phonemic and animal fluency, this association is strongest with animal fluency (Henry et al., 2004). Animal fluency is also more strongly associated than phonemic fluency with confrontation naming as evaluated by the Boston Naming Test (Henry & Crawford, 2004). Animal fluency has demonstrated sensitivity to age-associated cognitive decline (Tombaugh et al., 1999) head injury (Henry & Crawford, 2004), psychiatric disturbance (Henry & Crawford, 2005), cortical (Crossley et al., 1997) and subcortical (Henry et al., 2004) dementias. Education is strongly associated with animal fluency scores: as with phonemic fluency, older adults with the highest levels of educational have been shown to generate twice the number of animals as those with the lowest level of education (Crossley et al., 1997).

Demographic Variables

Participant age (in years), sex (male/female) and educational attainment (years of formal schooling) were obtained via self-report in the CSHA.

Measurement Model

The measurement model of the CSHA neuropsychological battery as identified by Tuokko and colleagues (2009) is displayed graphically in Figure 1. The investigators structured a confirmatory factor analysis (CFA) model based on Cattell-Horn-Carol theory and previous studies of the Wechsler Intelligence and Memory Scale (Bowden et al., 1997) and applied it to data from CSHA-1. The model comprised three factors: 1) Verbal Ability, upon which scores from Animal fluency, Phonemic fluency, WAIS Comprehension, WAIS Similarities and the Token test loaded, 2) Visuospatial Ability, upon which BVRT, WAIS Digit Symbol and WAIS Block Design loaded, and 3) Long-term Retention, upon which RAVLT, BVRT and WMS Information loaded.

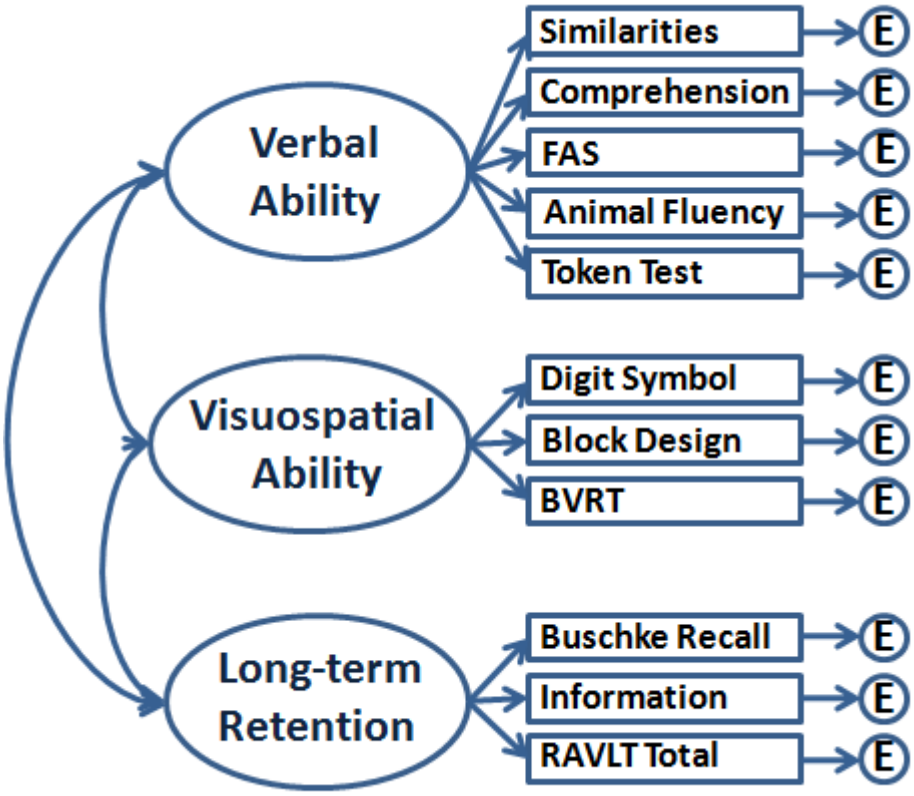

Figure 1. Measurement Model of the CSHA Neuropsychological Test Battery.

Following model identification the investigators initiated a series of post-hoc adjustments in order to optimize model fit. Taking a theory-based approach, Animal fluency was cross-loaded onto Long-term Retention in addition to its loading on Verbal Ability. Residuals were allowed to correlate between Phonemic and Animal fluency, RAVLT and Buschke total free recall, Buschke and WMS Information, RAVLT and Phonemic fluency, and between Buschke and Animal fluency. Fit indices for the baseline model and the final model as published by Tuokko and colleagues (2009) are presented in Table 1. For the purposes of the present analyses, this three-factor model was adopted without post-hoc modification.

Table 1. Fit of the Measurement Model to CSHA-1 Data (Tuokko et al., 2009).

Model	Sample	X^2	df	CFI	RMSEA (90% CI)	SRMR
Baseline Model	Exploratory	328.42	39	0.92	0.10 (0.09-0.11)	0.06
	Validation	297.60	39	0.93	0.10 (0.08-0.11)	0.05
Adjusted Model	Exploratory	134.11	35	0.97	0.06 (0.05-0.08)	0.03
	Validation	151.41	35	0.97	0.07 (0.06-0.08)	0.03

Note. Df = Degrees of frequency; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; CI = Confidence interval; SRMR = Standardized root mean residual.

Statistical Analyses

Data Preparation

Missing Data: The amount of missing neuropsychological test data was evaluated and missing data points were corrected for using regression imputation. Regression imputation estimates the values of missing data points for each case based on a linear combination of all other observed variables (Brown, 2006). Predicted values from the regression equations can then be used to replace missing values. Regression imputation is preferable to listwise deletion of cases with missing data because 1) CFA requires large sample sizes (though there is disagreement in the literature, a minimum of 10 participants per parameter in the measurement model is considered acceptable) in order to generate reliable estimates, thereby making deletion of cases undesirable and 2) Missing cognitive data in older adult populations can rarely be presumed missing completely at random (Anstey et al., 2001), and therefore removal of cases with missing data can introduce systematic bias into results, especially when more than 5% of cases include missing data.

Confirmatory Factory Analysis (CFA)

CFA was used to evaluate the factor structure of the CSHA test battery in the study sample. CFA is a special case of structural equation modeling (SEM) where directionality in relationships among factors is not specified. Thus CFA is strictly concerned with evaluating direct effects of latent factors on measured variables and covariances among factors. Like all forms of SEM, CFA is theory-driven and requires the investigator to define the model to be tested. This model is used to estimate a population covariance matrix which is compared to the observed matrix. The greater the similarity between estimated and observed covariance matrices, the better the model fit.

As illustrated by Widaman & Reise (1997), each measured variable in a CFA model can be represented as y_{ji} where *j* refers to the *j*th measured variable and *i* refers to the *i*th participant. Each y_{ji} is represented as the deviation of the raw score for person *i* from the mean of variable *j*. Each measured variable is represented as a linear function of its associated latent factor, η_k (*k* referring to the number of factors in the measurement model) and an error term, ϵ_{ji}. Similar to y_{ji}, the η_k value for each participant is represented as the difference between their η_k value and the mean value of η_k. Because both y_{ji} and η_k are represented as deviation scores, the relationship of a measured variable to a latent factor can be presented as a linear regression formula (with λ representing the regression of *y* on *η*):

$$y_{ji} = \lambda_{j1}\eta_{1i} + \lambda_{j2}\,\eta_{2i} + \ldots + \lambda_{ji}\eta_{mi} + \epsilon_{ji}$$

Thus, the measurement model identified by Tuokko and colleagues (2009), which has 11 measured variables and three latent factors would be represented as:

$$\begin{array}{lclcl}
y1i & = & \lambda 11 & & + \; \epsilon 1i \\
y2i & = & \lambda 21 & & + \; \epsilon 2i \\
y3i & = & \lambda 31 & & + \; \epsilon 3i \\
y4i & = & \lambda 41 & & + \; \epsilon 4i \\
y5i & = & \lambda 51 & \eta 1i & + \; \epsilon 5i \\
y6i & = & \lambda 62 & \eta 2i & + \; \epsilon 6i \\
y7i & = & \lambda 72 & \eta 3i & + \; \epsilon 7i \\
y8i & = & \lambda 83 & & + \; \epsilon 8i \\
y9i & = & \lambda 93 & & + \; \epsilon 9i \\
y10i & = & \lambda 103 & & + \; \epsilon 10i \\
y11i & = & \lambda 113 & & + \; \epsilon 11i
\end{array}$$

Given a sample of N participants with scores on each of the 11 observed variables, these equations must be redefined by expanding y, η and ϵ along i (participants). Specifically, y would become a (p x N) matrix of scores of the N participants on the p variables, η would become a (m x N) matrix of the scores of N participants on the m latent factors, and ϵ would become a (p x N) matrix of scores of the N participants on the p measurement residuals. By postmultiplying each side of this expanded equation by its transpose and dividing each side by N-1, the result is:

$$\Sigma = \Lambda\Psi\Lambda' + \theta_{\epsilon}$$

This formula describes a linear model of covariances among the p measured variables. In other words, this is a covariance structure model. Σ is a (p x p) matrix of covariances among the p measured variables, Λ is a (p x m) matrix of loadings of the p variables on the m latent factors, Ψ is an (m x m) matrix of covariances among the latent factor scores, and θ_{ϵ} is a (p x p) matrix of covariances among measurement residuals. For the model identified by Tuokko and colleagues (2009), this formula can be expanded to:

$$\begin{matrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} & \sigma_{17} & \sigma_{18} & \sigma_{19} & \sigma_{110} & \sigma_{111} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} & \sigma_{27} & \sigma_{28} & \sigma_{29} & \sigma_{210} & \sigma_{211} \\
\sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} & \sigma_{36} & \sigma_{37} & \sigma_{38} & \sigma_{39} & \sigma_{310} & \sigma_{311} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} & \sigma_{46} & \sigma_{47} & \sigma_{48} & \sigma_{49} & \sigma_{410} & \sigma_{411} \\
\sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} & \sigma_{56} & \sigma_{57} & \sigma_{58} & \sigma_{59} & \sigma_{510} & \sigma_{511} \\
\sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_{66} & \sigma_{67} & \sigma_{68} & \sigma_{69} & \sigma_{610} & \sigma_{611} \\
\sigma_{71} & \sigma_{72} & \sigma_{73} & \sigma_{74} & \sigma_{75} & \sigma_{76} & \sigma_{77} & \sigma_{78} & \sigma_{79} & \sigma_{710} & \sigma_{711} \\
\sigma_{81} & \sigma_{82} & \sigma_{83} & \sigma_{84} & \sigma_{85} & \sigma_{86} & \sigma_{87} & \sigma_{88} & \sigma_{89} & \sigma_{810} & \sigma_{811} \\
\sigma_{91} & \sigma_{92} & \sigma_{93} & \sigma_{94} & \sigma_{95} & \sigma_{96} & \sigma_{97} & \sigma_{98} & \sigma_{99} & \sigma_{910} & \sigma_{911} \\
\sigma_{101} & \sigma_{102} & \sigma_{103} & \sigma_{104} & \sigma_{105} & \sigma_{106} & \sigma_{107} & \sigma_{108} & \sigma_{109} & \sigma_{1010} & \sigma_{1011} \\
\sigma_{111} & \sigma_{112} & \sigma_{113} & \sigma_{114} & \sigma_{115} & \sigma_{116} & \sigma_{117} & \sigma_{118} & \sigma_{119} & \sigma_{1110} & \sigma_{1111}
\end{matrix}$$

$$=$$

$$\begin{matrix}
\lambda_{11} & 0 & 0 \\
\lambda_{21} & 0 & 0 \\
\lambda_{31} & 0 & 0 \\
\lambda_{41} & 0 & 0 \\
\lambda_{51} & 0 & 0 \\
0 & \lambda_{12} & 0 \\
0 & \lambda_{22} & 0 \\
0 & \lambda_{32} & 0 \\
0 & 0 & \lambda_{13} \\
0 & 0 & \lambda_{23} \\
0 & 0 & \lambda_{33}
\end{matrix}
\begin{matrix}
\Psi_{11} & \Psi_{12} & \Psi_{13} & \lambda_{11} & \lambda_{21} & \lambda_{31} & \lambda_{41} & \lambda_{51} & 0 & 0 & 0 & 0 & 0 & 0 \\
\Psi_{21} & \Psi_{22} & \Psi_{23} & 0 & 0 & 0 & 0 & 0 & \lambda_{12} & \lambda_{22} & \lambda_{32} & 0 & 0 & 0 \\
\Psi_{31} & \Psi_{23} & \Psi_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{13} & \lambda_{23} & \lambda_{33}
\end{matrix}$$

$$+$$

$$\begin{matrix}
\theta_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \theta_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \theta_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \theta_{44} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \theta_{55} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \theta_{66} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \theta_{77} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{88} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{99} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{1010} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{1111}
\end{matrix}$$

As can be seen from the diagonal matrix of residual covariances, this model assumes uncorrelated measurement residuals. This is enforced by constraining to zero the covariances of the eleven measurement residuals in the model. If this model were fit to a covariance matrix from an independent sample of *N* participants, the equation would be:

$$S \cong \hat{\Lambda}\hat{\Psi}\hat{\Lambda}' + \hat{\theta}_{\epsilon} = \Sigma_i$$

In this equation S is an observed $p \times p$ sample covariance matrix, while $\hat{\Lambda}$, $\hat{\Psi}$, $\hat{\theta}_{\epsilon}$ and Σ_i represent estimates of population parameters in this sample. The observed covariances in S are estimated by the CFA formula $\Lambda\Psi\Lambda' + \theta_{\epsilon}$, which produces an estimate of Σ (values for covariances among the measured variables under the assumption that the model accurately represents S). Model fit can then be determined based on the fit of $\hat{\Lambda}\hat{\Psi}\hat{\Lambda}' + \hat{\theta}_{\epsilon}$ to S relative to Σ.

<u>Model Identification:</u> The Maximum Likelihood (ML) method is the most common method of parameter estimation for CFA models and was used in the present analyses. The ML method estimates parameters based on the likelihood function $L(w|y)$, which tests the likelihood of parameter w given the observed data y. Thus, the ML method identifies the probability distribution that makes the observed data most likely. When variables are continuous and approximate a normal distribution, the ML method has been shown to yield optimal parameter estimates (Joreskog & Sorbom, 1988). ML estimation also yields a likelihood ratio X^2 statistic that tests whether the covariance matrices reproduced from the estimated parameters differ significantly from the observed sample covariance matrices. As will be discussed, the analyzed relational indices for the present analyses were covariance matrices due to the use of invariance testing.

Normality. An important assumption of the ML method is univariate and multivariate normality. To satisfy this assumption, the univariate skewness and kurtosis estimates of each of the 11 observed variables (tests) were examined. Absolute univariate normality

requires that a variable have skewness and kurtosis values of 0. However, absolute normality is rare and the ML method is resilient to deviations from normality that fall within the mild to moderate range (Benson, 1994). In fact, the ML method can even tolerate extreme violation of normality if it is observed in less than 50% of the indicators in a measurement model (Benson, 1994). For the present analyses absolute values for skewness and kurtosis that are less than 2.0 were considered acceptable deviations from normality. Multivariate normality was evaluated using the multivariate kurtosis estimate provided by Mardia's coefficient (Mardia, 1970). In order to satisfy the multivariate normality assumptions of the ML method, Mardia's coefficient must be 10.00 or lower.

Fit Indices: *Chi-Square Value (X^2).* X^2 is an inferential statistic that tests the null hypothesis that the measurement model fits the analyzed covariance matrix perfectly (Loehlin, 1988). Because the *T*-statistic from which X^2 is derived is calculated by multiplying *N*-1 (sample size minus one) by the minimum estimate of the fit function (e.g., ML), increasing sample size inflates *T* without affecting the referent X^2 distribution against which the significance of the model is tested. As a result, the probability of rejecting the null hypothesis can be inflated or deflated in cases where sample sizes are very large or small, rendering X^2 an unreliable fit index under these circumstances. Given its limitations, it is important to consider other fit indices when evaluating a model.

$$T = (N-1)F_{min}$$

Relative Chi-Square. The relative chi-square calculates a parsimony index by dividing the X^2 value of a model by its degrees of freedom (Bollen, 1989). In this manner, model fit (as estimated by its X^2 value) is adjusted based on model complexity (e.g., the number of specified parameters. The criterion for acceptance ranges from less than 2 to less than 5.

$$X^2/df$$

Comparative Fit Index (CFI). The CFI provides a measure of model fit by comparing the chosen measurement model with an "independence" model where variables are assumed to be uncorrelated (Bentler, 1990). The CFI is ideal for use in analyses of smaller samples because its estimates are not influenced by sample size. CFI should be 0.90 or higher in order for a model to be acceptable and values above 0.95 are needed in order to establish optimal model fit (Hu & Bentler, 1999).

$$CFI = [D(\text{Independence Model}) - D(\text{Chosen Model})]/ D(\text{Independence Model})$$

Where $D = X^2 - df$

Root mean squared error of approximation (RMSEA) (Steiger, 1990). RMSEA provides a parsimony-adjusted "badness of fit" estimate based on the square root of the average of covariance residuals. Estimates of 0.05 and lower are considered optimal, estimates between 0.06 and 0.08 are acceptable and those over 0.09 are poor (Byrne, 2001).

$$RMSEA = \sqrt{([X^2/df] - 1)/(N - 1)}$$

Invariance Testing

Invariance testing was conducted in accordance with the recommendations of Brown (2006), and involved placing progressively stricter constraints on the measurement model and fitting these adjusted models to the LE and HE sample data.

Model Identification. Invariance testing requires modification of the equations for CFA to allow for multiple-group linear modeling. Specifically, a subscript (*g*) is added to all parameters to denote group membership:

$$S_g \cong \hat{\Lambda}_g \hat{\Psi}_g \hat{\Lambda}_g' + \hat{\theta}_{g\epsilon} = \Sigma_{gi}$$

Group comparisons can only be conducted using covariance matrices. Covariance structure models like the one above produce inaccurate group comparisons when applied to correlation matrices because, unlike covariance matrices, parameters derived from within-group correlation matrices are not standardized (or on the same scale) in a way that permits group comparison.

Configural Invariance. In multiple-group analyses a single X^2 value can be obtained that assesses aggregate fit across the estimated and observed values for *Sg*. The fit of this aggregate model is often used as a basis for establishing configural invariance. However aggregate models can be misleading because good fit estimates can be observed even in the presence of significant between-group discrepancies in model fit. Thus it is considered best practice when evaluating configural invariance to run single-group models on each of the observed groups and consider the fit of each of these single-group models in conjunction with the aggregate model. For the present analyses, configural

invariance will be tested by 1) fitting the measurement model to the LE and HE samples independently and 2) examining the fit of the aggregate model. The fit of nested models were subsequently tested against the fit of this aggregate model.

Metric Invariance. In multigroup analyses parameters can be constrained equal across groups simply by removing the *g* subscript from the chosen parameter. For example, the formula for testing the invariance of model factor loadings would be:

$$S_g = \hat{\Lambda}\hat{\Psi}_g\hat{\Lambda}' + \hat{\theta}_{g\epsilon} = \Sigma_{gi}$$

In this formula, the subscripts for Λ (the factor loading covariance matrix) are released and thus the same factor loading values are applied to all groups. X^2 difference tests are then used to evaluate the fit of this constrained (or nested) model relative to the baseline model where all parameters are free to differ between groups. Difference tests are conducted by obtaining the difference between the aggregate X^2 values and degrees of freedom of a baseline model and nested model as follows:

Model	X^2	*df*
Baseline	*100*	*20*
Nested	*- 97*	*- 18*
Difference	*3*	*2*

The significance of the difference in X^2 value between the baseline and nested models (with an associated *df* equal to the difference between the *df* of the baseline and nested model) can then be examined using a X^2 table. Continuing with the example

above, a X^2 value of 3.00 with a *df* of 2 is not statistically significant ($p = 0.22$). This means that constraining factor loadings across groups does not significantly impact model fit, and thus the factor loadings can be considered invariant across groups. It is subsequently possible to continue invariance testing by constraining additional model parameters (e.g., $\hat{\Psi}$). In cases where X^2 differences are statistically significant it is not possible to consider the constrained parameters invariant across groups. When parameters are non-invariant, it is always the baseline model that provides better fit (e.g., significant differences never mean that the constrained model is a *better* representation of the data than the baseline model).

Partial invariance testing can be pursued when there is a statistically significant difference between the fit of baseline and nested models. This represents a simple extension of invariance testing where constraints are placed on only some of the values of a covariance matrix. For example, the expanded formula for $\hat{\Lambda}_g$ is:

$$
\begin{matrix}
\lambda_{11g} & 0 & 0 \\
\lambda_{21g} & 0 & 0 \\
\lambda_{31g} & 0 & 0 \\
\lambda_{41g} & 0 & 0 \\
\lambda_{51g} & 0 & 0 \\
0 & \lambda_{12g} & 0 \\
0 & \lambda_{22g} & 0 \\
0 & \lambda_{32g} & 0 \\
0 & 0 & \lambda_{13g} \\
0 & 0 & \lambda_{23g} \\
0 & 0 & \lambda_{33g}
\end{matrix}
$$

The values for each factor loading in the example above are estimated separately in each of the *g* samples that the model is being fit to. To test the factorial invariance of this model, the *g* subtext would be removed, and the formula would be:

$$\begin{matrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ \lambda_{41} & 0 & 0 \\ \lambda_{51} & 0 & 0 \\ 0 & \lambda_{12} & 0 \\ 0 & \lambda_{22} & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & 0 & \lambda_{13} \\ 0 & 0 & \lambda_{23} \\ 0 & 0 & \lambda_{33} \end{matrix}$$

If the fit of these two models is significantly different then full factorial invariance cannot be established. In these cases it is possible to test the invariance of each individual factor by releasing constraints on its loadings while retaining the constraints on the loadings for the other factors:

$$\begin{matrix} \lambda_{11g} & 0 & 0 \\ \lambda_{21g} & 0 & 0 \\ \lambda_{31g} & 0 & 0 \\ \lambda_{41g} & 0 & 0 \\ \lambda_{51g} & 0 & 0 \\ 0 & \lambda_{12} & 0 \\ 0 & \lambda_{22} & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & 0 & \lambda_{13} \\ 0 & 0 & \lambda_{23} \\ 0 & 0 & \lambda_{33} \end{matrix}$$

In this example above, factors 2 and 3 are constrained equal, but factor 1 is free to differ between groups. Corresponding X^2 difference values will reveal whether any single factor is driving non-invariance. If so, it is possible to identify the specific tests that are driving non-invariance by systematically removing constraints on the individual tests that load onto the non-invariant factor and examine corresponding changes in X^2 values. For the present analyses metric invariance will be tested by 1) examining the invariance of factor loadings (Λ) across HE and LE samples, 2) pursuing partial invariance testing if factor loadings cannot be constrained equal across groups, and 3) adding constraints on factor covariances (Ψ) if full or partial invariance of the factor loadings is established.

Scalar Invariance. Scalar invariance would only be examined if test scores did not significantly differ across HE and LE samples. Examining scalar equivalence is redundant when mean-level group differences are known to exist because it involves constraining mean values of each indicator equal across groups. Education is known to impact cognitive performance and thus performance on each of the eleven observed tests is expected to differ significantly between the HE and LE sample.

Results

Participant Characteristics

Of 1,466 participants who underwent a physician's examination and neuropsychological assessment at CSHA-2, 1,001 (68%) identified English as their native language. Five hundred five English-speaking participants (34% of the CSHA-2 assessment cohort) were determined by consensus diagnosis to have NCI and thus met all inclusion criteria for the present study. Table 2 presents the demographic characteristics of the study sample (n = 505) vs. the entire CSHA-2 assessment cohort (including French speaking participants and those with a consensus diagnosis other than NCI; n = 1,466).

Table 2. Characteristics of the study sample vs. the full CSHA-2 cohort.

Sample	Study	Cohort	Sample vs. Cohort*
Characteristic			
n	505	1466	
Age, mean (SD)	81.37 (6.33)	82.03 (6.69)	7.39 (0.01)
Education, mean (SD)	11.24 (3.80)	9.81 (3.99)	91.26 (<0.001)
Sex, count (%) Female	275 (54.46)	859 (58.59)	5.43 (0.02)
3MS, mean (SD)	87.93 (9.16)	79.69 (12.78)	256.32 (<0.001)

Note. SD = standard deviation; 3MS = Modified Mini Mental Status Examination; CSHA = Canadian Study of Health and Aging; * values denote X^2 (*p*-value)

As expected, participants eligible for inclusion in this analysis were significantly younger and had a higher level of education than the rest of the cohort. The study sample also had a significantly higher proportion of men and significantly higher 3MS scores than the full CSHA-2 assessment cohort.

Demographic differences between the study sample and the full CSHA-2 cohort can be attributed to the fact that participants in the study sample were determined by consensus diagnosis to have NCI, whereas the full cohort included cognitive impairment, dementia and other neurological disorders. Age, low education and female sex are all risk factors for cognitive impairment (Yen et al., 2004) and thus it is not surprising that the full cohort differed from the (cognitively intact) study sample along these dimensions.

Of the 505 participants included in the study sample, 120 reported fewer than nine years of education and were included in the LE sample. The remaining 385 reported nine or more years of education and comprised the HE sample. The distribution of years of education in the HE and LE samples are illustrated in Figure 2 and Figure 3.

Table 3 presents the demographic characteristics and raw test scores of the LE and HE samples. Relative to the HE sample, the LE sample was significantly older, included a higher proportion of males, had less education and lower 3MS scores. As expected based on the well-documented association between education and cognitive performance (McCarthy et al., 2003), the HE sample had higher scores on all eleven neuropsychological tests.

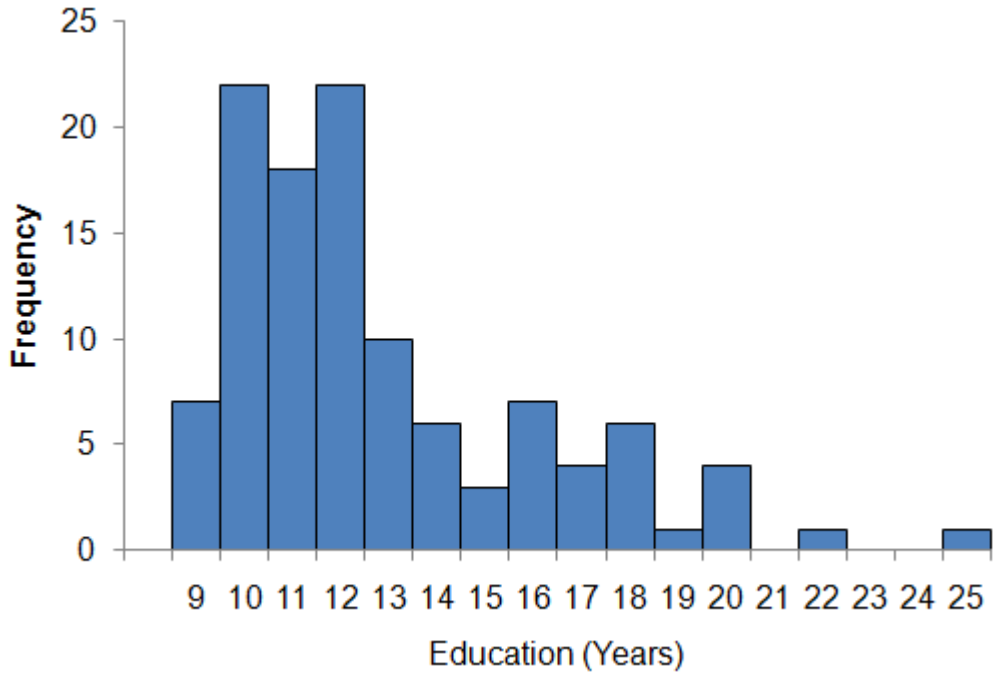

Figure 2. Distribution of years of formal schooling in the HE sample.

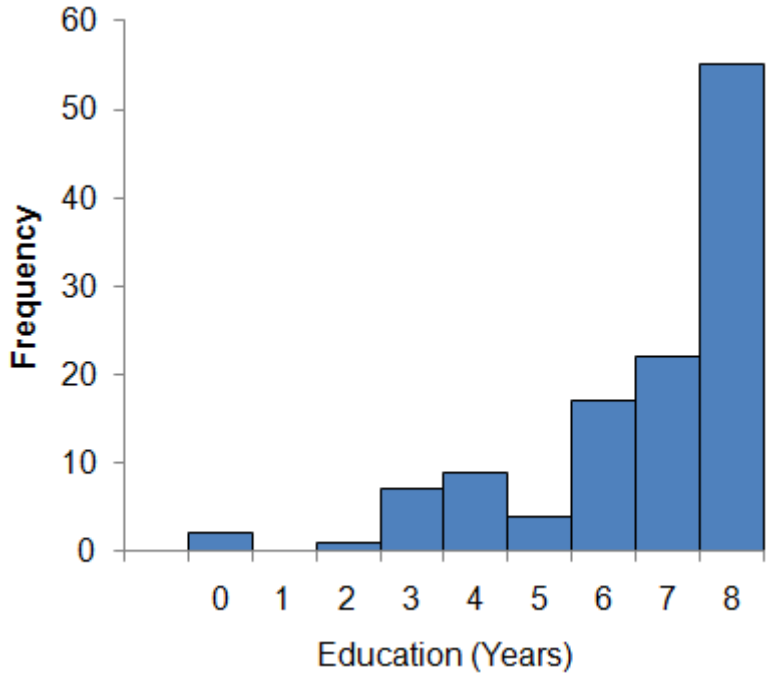

Figure 3. Distribution of years of formal schooling in the LE sample.

Table 3. Demographic characteristics and mean test scores of the study sample.

Sample	LE (<9 Years)	HE (>9 Years)	LE vs. HE*
Characteristic			
n	120	385	
Age, mean (SD)	82.85 (6.45)	80.91 (6.24)	8.40 (0.003)
Education, mean (SD)	6.55 (1.90)	12.66 (3.01)	19.24 (<0.001)
Sex, count (%) Female	46 (38.33)	229 (59.48)	16.04 (<0.001)
3MS, mean (SD)	81.69 (10.79)	89.95 (7.56)	57.33 (<0.001)
Observed variable, mean (SD)			
Benton Visual Retention	10.83 (2.58)	12.42 (1.99)	35.79 (<0.001)
Total Buschke Cued Recall	25.17 (4.61)	26.22 (4.13)	5.76 (0.02)
Rey AVLT	33.52 (9.14)	40.43 (9.82)	32.38 (<0.001)
Token Test	36.67 (5.96)	40.93 (4.39)	31.26 (<0.001)
Verbal fluency, Phonemic	23.56 (10.95)	32.74 (12.04)	42.07 (<0.001)
Verbal fluency, Animal	13.42 (3.93)	15.54 (4.29)	21.41 (<0.001)
WAIS Block Design	8.71 (4.15)	10.40 (4.29)	12.65 (<0.001)
WAIS Comprehension	8.30 (2.90)	10.53 (2.99)	42.33 (<0.001)
WAIS Digit Symbol	22.54 (8.62)	32.03 (10.12)	53.67 (<0.001)
WAIS Similarities	5.43 (3.82)	9.19 (3.17)	70.44 (<0.001)
WMS Information	4.72 (1.31)	9.19 (3.57)	20.74 (<0.001)

Note. LE = lower-educated sample; HE = higher-educated sample; SD = standard deviation; 3MS = Modified Mini Mental Status Examination; Rey AVLT = Rey Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales; * values denote X^2 (*p*-value)

Data Preparation

Missing Data: Four percent of data points in the LE sample and 2% of data in the HE sample were missing and thus these values were imputed. Although missing data in both samples was less than 5% and removal of cases with missing data would have been acceptable, imputation was considered preferable because the proportion of missing data in the LE sample was double that of the HE sample and therefore deletion of cases with missing data could introduce bias into subsequent group comparisons.

Matched Sampling: Because the HE and LE samples differed in their age and sex distributions, matched sampling was implemented to ensure that study findings were not an artefact of these between-group differences. To evaluate the impact of demographic differences on model fit and measurement invariance across the HE and LE samples, all analyses were conducted first on the full study sample, and subsequently on the matched subsamples. Subsamples were extracted from the full study sample using SPSS software. The syntax that was used to extract matched subsamples was modified from existing syntax that is available via the following web link: http://www.spsstools.net/Syntax/RandomSampling/findRandomPairsOfCasesWithSameCharacteristics.txt. This syntax used the LE sample as "cases" and matched each LE "case" with a randomly selected HE "control" of the exact same age or sex. Because age is a continuous variable it would have been possible to use "minimum distance" matching to assign LE participants the HE counterpart that is closest to them in age (but not necessarily the same age). However sex is dichotomous and permits only direct matching. In the interest of continuity and because the HE sample was of a sufficient size to make this feasible, direct matching was used to extract both the age- and sex-matched subsamples.

The demographic characteristics of the age-matched and sex-matched samples are presented in Table 4a and 4b, respectively. These tables illustrate that the demographic characteristics of the matched samples closely resemble the characteristics of the full study sample (aside from the dimension on which they were matched).

Normality Assumptions: The univariate characteristics and Mardia coefficient of the 11 observed tests in the full HE and LE samples are presented in Table 5. See Appendix A for these characteristics in the age-and sex-matched samples. See Appendix B for the bivariate associations among the eleven tests in all samples. Mardia's coefficient was below 10 in all samples, indicating that the multivariate normality assumptions of the ML method were satisfied. Skewness of the distribution of test scores consistently fell below 2.0. Kurtosis estimates for the Token Test were above the 2.0 criterion for normality across all samples. Kurtosis estimates for WMS information were also slightly above 2.0. Because only one of the eleven indicators seriously violates normality, it is unlikely that ML estimates for the model will be seriously biased (Benson & Fleishman, 1994).

Table 4a. Characteristics of the age-matched samples.

Sample	LE (<9 Years)	HE (>9 Years)
Characteristic		
n	115	115
Age, mean (SD)	82.36 (6.10)	82.36 (6.10)
Education, mean (SD)	6.62 (1.86)	12.79 (3.20)
Sex, count (%) Female	43 (37.40)	67 (58.30)
3MS, mean (SD)	81.99 (10.46)	89.10 (8.54)

Note. LE = lower-educated; HE = higher-educated sample; SD = standard deviation; 3MS = Modified Mini Mental Status Exam.

Table 4b. Characteristics of the sex-matched samples.

Sample	LE (<9 Years)	HE (>9 Years)
Characteristic		
n	120	120
Age, mean (SD)	82.85 (6.45)	81.04 (6.51)
Education, mean (SD)	6.55 (1.91)	12.80 (3.33)
Sex, count (%) Female	46 (38.30)	46 (38.30)
3MS, mean (SD)	81.69 (10.79)	89.68 (7.78)

Note. LE = lower-educated sample; HE = higher-educated sample; SD = standard deviation; 3MS = Modified Mini Mental Status Examination.

Table 5. Univariate and multivariate characteristics of the 11 observed variables in the full study sample.

Sample	LE Sample				HE Sample			
Variable	Min	Max	Skew.	Kurt.	Min	Max	Skew.	Kurt.
WMS Information	1	6	-0.91	-0.15	1	6	-1.51	2.03
Buschke Recall	13	35	-0.17	-0.11	11	35	-0.47	0.13
Rey AVLT	13	63	0.69	0.58	17	69	0.28	-0.27
Digit Symbol	5	60	0.71	1.69	8	62	0.33	-0.16
Block Design	0	20	0.27	0.08	0	25	0.46	0.62
BVRT	2	16	-0.62	0.18	3	16	-0.77	1.31
Phonemic Fluency	3	55	0.79	0.43	7	67	0.24	-0.52
Comprehension	2	15	-0.14	-0.54	2	16	-0.48	0.07
Token Test	16	66	1.01	3.76	18	66	1.08	6.58
Animal Fluency	6	25	0.64	0.24	4	28	0.11	-0.22
Similarities	0	14	0.29	-1.13	0	14	-0.64	-0.62
Mardia				6.50				9.02

Note. LE = lower-educated; HE = higher-educated; WMS = Weschler Memory Scales; AVLT = Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales.

Fit of the Measurement Model to CSHA-2 Neuropsychological Test Scores

To confirm the fit of the measurement the model identified by Tuokko and colleagues (2009) in the current study sample and establish its configural invariance, the model was first fit independently to the HE and LE samples. An aggregate model was subsequently fit across both samples. The fit of these models in the full study sample are presented in Table 6. The fit of these models in the age-and sex-matched samples are presented in Appendix C. Squared multiple correlations and regression coefficients for the model in the HE and LE samples are presented in Table 7. These characteristics in the matched samples are presented in Appendix D. Confidence intervals and standard errors (for unstandardized regression weights) were obtained via 1,000 bootstrap iterations.

Configural invariance: Baseline model fit in the full LE and HE samples.

Sample	X^2	*df*	Pars.	CFI	RMSEA (90% CI)
Aggregate sample (n=505)	181.60	82	2.22	0.93	0.05 (0.04-0.06)
HE sample only (n=385)	137.57	41	3.36	0.91	0.08 (0.06-0.09)
LE sample only (n=120)	44.02	41	1.07	0.99	0.03 (0.00-0.07)

Note. LE = lower-educated; HE = higher-educated; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval.

<u>HE Sample.</u> The measurement model demonstrated adequate fit to the test scores of the 385 participants with 9 or more years of education. All parameter estimates were significant and the standardized estimates of factor loadings all had values greater than 0.30, qualifying them as “salient” loadings (Brown, 2006). Estimates of model parsimony and fit were within acceptable limits. However, the obtained RMSEA value of 0.08 was on the upper margin of acceptable limits (Brown, 2006), suggesting that model fit was

suboptimal. Similarly, the obtained CFI value of 0.91 was close to the minimum acceptable value of 0.90. Considering together the parsimony index and two fit indices, it can be concluded that the measurement model provided an adequate but suboptimal representation of the psychological constructs measured by the CSHA neuropsychological test battery in the HE sample. Fit of the measurement model to the age-and sex-matched HE subsamples was very similar to fit in the full HE sample.

LE Sample. The measurement model demonstrated excellent fit to the test scores of the 120 participants with eight of fewer years of formal schooling. The X^2 value, parsimony index and fit indices all indicate a highly parsimonious model demonstrating exceptionally good fit. All parameter estimates were statistically significant and standardized estimates for all tests except WMS Information were salient ($\geq$0.30). Based on these estimates the measurement model provides a strong representation of the psychological constructs measured by the CSHA neuropsychological test battery in the LE sample. The model demonstrated similar fit to the age-and sex-matched subsamples.

Aggregate Sample. The measurement model demonstrated good fit to the test scores of the aggregate study sample of 505 participants. The parsimony index, CFI and RMSEA all suggested that the model was parsimonious and well-fit, representing a workable model of the psychological constructs measured by the CSHA neuropsychological test battery. These criteria establish the configural invariance of the measurement model across HE and LE sample and allowed for subsequent examination of metric invariance. The fit of the measurement model to the age-and sex-matched subsamples was very similar to its fit to the full sample, further establishing its invariance in these subsamples.

Table 7. Squared multiple correlations, standardized and unstandardized coefficients for the model when applied to the full HE and LE samples.

Construct	Variable	Sample	SMC (95% CI)	B (95% CI)	β (95% CI)	SE
Verbal Ability	Phonemic Fluency	HE	0.42 (0.33-0.53)	4.03 (3.28-5.32)	0.65 (0.58-0.73)	0.04
		LE	0.60 (0.42-0.75)	5.26 (3.65-7.80)	0.77 (0.65-0.86)	0.05
	Animal Fluency	HE	0.43 (0.34-0.51)	0.36 (0.31-0.41)	0.66 (0.59-0.72)	0.04
		LE	0.26 (0.13-0.42)	0.24 (0.16-0.37)	0.51 (0.36-0.65)	0.07
	Token Test	HE	0.13 (0.05-0.20)	0.30 (0.17-0.39)	0.35 (0.23-0.45)	0.05
		LE	0.28 (0.11-0.48)	0.50 (0.31-0.76)	0.53 (0.34-0.69)	0.09
	Comprehension	HE	0.42 (0.30-0.51)	0.25 (0.19-0.31)	0.65 (0.55-0.72)	0.04
		LE	0.31 (0.16-0.47)	0.19 (0.13-0.27)	0.55 (0.40-0.68)	0.07
	Similarities	HE	0.47 (0.36-0.56)	0.31 (0.25-0.38)	0.69 (0.60-0.75)	0.04
		LE	0.56 (0.39-0.71)	0.34 (0.27-0.44)	0.75 (0.62-0.84)	0.05
Visuospatial Ability	Digit Symbol	HE	0.61 (0.50-0.71)	3.20 (2.62-3.94)	0.78 (0.71-0.84)	0.04
		LE	0.63 (0.46-0.77)	2.88 (2.18-4.46)	0.79 (0.68-0.88)	0.05
	Block Design	HE	0.61 (0.50-0.71)	0.31 (0.25-0.38)	0.57 (0.50-0.66)	0.04
		LE	0.33 (0.17-0.48)	0.35 (0.23-0.46)	0.57 (0.41-0.69)	0.07
	BVRT	HE	0.31 (0.21-0.42)	0.14 (0.11-0.18)	0.56 (0.46-0.65)	0.05
		LE	0.34 (0.21-0.50)	0.22 (0.17-0.32)	0.59 (0.46-0.71)	0.06
Long-term Retention	Cued Recall	HE	0.38 (0.26-0.51)	0.33 (0.26-0.45)	0.62 (0.51-0.71)	0.05
		LE	0.35 (0.04-0.52)	0.31 (0.03-0.45)	0.59 (0.21-0.72)	0.11
	Information	HE	0.11 (0.03-0.22)	0.04 (0.02-0.07)	0.33 (0.19-0.47)	0.07
		LE	0.02 (0.00-0.12)	0.02 (0.01-0.06)	0.14 (0.10-0.35)	0.13
	Rey AVLT	HE	0.61 (0.45-0.77)	3.00 (2.25-3.84)	0.78 (0.67-0.88)	0.05
		LE	0.94 (0.67-3.59)	3.25 (2.31-11.4)	0.97 (0.82-1.89)	0.36

Note. LE = lower-educated; HE = higher-educated; AVLT = Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales; SMC = squared multiple correlations; β = standardized regression weights; B = unstandardized regression weights; CI = confidence intervals.

Metric Invariance of the Measurement Model across HE and LE Samples

Metric invariance of the measurement model was tested as described in the Methods. The fit of the baseline model and nested models to the study sample are presented in Table 7. The difference X^2 tests that were used to establish measurement invariance are presented in Table 8.

Table 8. Fit of the baseline and nested models to the study sample.

Model	X^2	*df*	Pars.	CFI	RMSEA (90% CI)
Unconstrained	181.60	82	2.22	0.93	0.05 (0.04-0.06)
Constrained factors:					
3 (All Factors)	143.54	90	1.60	0.93	0.05 (0.03-0.06)
2 (Visuospatial Ability freed)	140.01	88	1.59	0.93	0.05 (0.03-0.07)
2 (Long-term Retention freed)	141.03	88	1.60	0.93	0.05 (0.03-0.07)
2 (Verbal Ability freed)	130.79	86	1.52	0.94	0.05 (0.03-0.06)

Note. LE = lower-educated; HE = higher-educated; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval;

Constraint of the loadings of the 11 observed variables on the three latent factors across the HE and LE samples resulted in a significant loss of fit relative to the baseline model (Table 9, Model A). Partial measurement invariance of the factor loadings was subsequently tested for each of the three latent factors. Release of the Long-term Retention (Table 9, Model B) and Visuospatial Ability (Table 9, Model C) factors did not appreciably improve model fit, but release of the constraints on the Verbal Ability factor rendered nonsignificant the differences in fit between the baseline and nested models (Table 9, Model D). With the Verbal Ability factor loadings identified as the

source of non-invariance, each test that loaded onto this factor was subsequently tested for invariance. With all other factor loadings in the measurement model constrained equal across the HE and LE samples, release of the loadings of WAIS Similarities (Table 9, Model E), WAIS Comprehension (Table 9, Model F) and Phonemic fluency (Table 9, Model G) did not result in an invariant measurement model. However, models releasing constraints on Animal fluency (Table 9, Model H) and the Token test (Table 9, Model I) did not differ significantly from the baseline model in terms of fit and were invariant.

Table 9. Metric invariance of the model across LE and HE samples.

Model	X^2	*df*	*p*-value	Invariant?
Metric invariance				
A. All factors constrained	20.67	8	0.01	No
B. Long-term Retention released	17.96	6	0.04	No
C. Visuospatial Ability released	14.95	6	0.02	No
D. Verbal Ability released	5.38	4	0.25	**Yes**
Partial metric invariance: Verbal Ability				
E. Similarities released	17.20	7	0.02	No
F. Comprehension released	16.41	7	0.02	No
G. Phonemic fluency released	18.15	7	0.02	No
H. Animal fluency released	13.71	7	0.06	**Yes**
I. Token test released	12.99	7	0.07	**Yes**
Partial invariance: All factors constrained*				
J. Factor loadings only	8.76	6	0.19	**Yes**
K. Factor loadings + Structural covariances	11.41	12	0.49	**Yes**

Note. LE = lower-educated; HE = higher-educated; df = degrees of freedom * non-invariant tests released.

Following release of the constraints on factor loadings for Animal fluency and the Token test, factorial invariance was achieved (Table 8, Model J). The subsequent addition of constraints on factor covariances across HE and LE samples did not impact the fit of this model or compromise its invariance (Table 9, Model K). These findings were replicated in the age-and sex-matched subsamples (Appendix E). Because the HE sample yielded significantly higher scores on all eleven tests included in the measurement model it was considered redundant to proceed with scalar invariance testing.

Discussion

Education is known to mediate the structure and function of the aging brain. Evidence from animal models suggests that cognitive engagement early in life increases hippocampal neurogenesis, axon arbour growth and synaptodendritic development (Lazarov et al., 2005). Illustrating the health benefits of cognitive engagement, these brain changes ameliorate amyloid deposition in mouse models of Alzheimer disease and foster resilience to behavioural expression of neuropathology (Mirochnic et al., 2009). As reviewed in the Introduction, education in humans similarly mediates age-associated microstructural changes in the integrity of the hippocampal formation (Piras et al., 2011). Education has also been shown behaviourally to ameliorate age-associated cognitive decline and delay the onset of dementia (Stern, 2009). Functionally, education and other markers of cognitive reserve mediate the neural networks that are activated during cognitive task performance with a trend toward increased recruitment of frontal regions and decreased reliance on parietal structures (Springer et al., 2005). Some studies (but not all; Springer et al., 2005) have linked education-associated patterns of activation with superior cognitive performance (Angel et al., 2010). These patterns of functional activation are thought to represent compensatory networks that allow educated older adults to remain cognitively intact longer than lower-educated counterparts (Stern, 2009).

Clinically the impact of education on cognitive ability has long been recognized and education-stratified normative data are available for many neuropsychological test batteries (Lezak, 2004). However, there has been little reflection on why neuropsychological tests elicit higher scores from higher vs. lower-educated individuals. There may be an implicit assumption that test scores reflect real, static baseline

differences in the abilities that are measured by neuropsychological tests. This could be a reasonable assumption, particularly for tests that measure verbal ability, crystallized knowledge or other abilities that are usually acquired from formal schooling. However, education has an almost global effect on neuropsychological test performance (Ivnik et al., 1996). For example, the Benton Visual Retention Test measures nonverbal recognition memory and facilitates no verbally-mediated encoding or retrieval strategies (Le Carret et al., 2003b). It is difficult to conceptualize the abilities acquired from education that would directly translate to performance on tests of nonverbal recognition memory, but education strongly influenced performance on the Benton Visual Retention Test in the present study sample. The assumption that test scores reflect real baseline differences in the abilities that are measured by neuropsychological tests also does not account for the experimental literature reviewed in this thesis that supports the presence of education-associated differences in the way cognitive tasks are processed and the neural correlates of task performance. If this assumption is unfounded then the conclusions that can be drawn from cross-group comparisons of neuropsychological test scores are severely limited. Although tests may retain their ecological validity, inferences regarding group differences in brain function that are made on the basis of non-invariant neuropsychological test scores would be problematic.

Invariance testing literature characterizes group differences in the mean value of scores on an observed variable (or test) as scalar non-invariance (Brown, 2006). Thus, the scalar non-invariance of neuropsychological tests across education-stratified samples has long been recognized, albeit using a different lexicon. Given the hierarchical nature of invariance testing, whereby scalar invariance can only be achieved under the assumption

of metric invariance and metric invariance can only be established under the assumption of configural invariance, an examination of the configural and metric equivalence of neuropsychological tests across low vs. higher levels of education was warranted in order to truly understand the impact of education on neuropsychological test performance. Thus, the objective of this thesis was to determine whether documented education-associated differences in the cognitive strategies and neural networks that are recruited during neuropsychological test performance are of a sufficient magnitude to compromise the configural, metric and scalar measurement equivalence of a battery of standardized neuropsychological tests.

Model Selection

CFA with invariance testing was selected as the framework for evaluating measurement equivalence. CFA is used widely as a way of evaluating measurement equivalence but it comes with assumptions that must be considered before reliable conclusions can be drawn. For example, CFA assumes that the specified measurement model provides a valid and empirically-supported conceptualization of the latent factors that are measured by the observed variables. It is important that the measurement model provides a conceptualization of the observed variables (or tests) that is consistent with the way scores from these tests are interpreted. A perfectly-fitting model that does not conceptualize the observed variables in a way that is consistent with prior research, theory and practice has no ecological validity and its invariance is of little consequence because results will not be generalizable. On the other hand, a specified model that corresponds perfectly with prior research, psychometric theory and practice but yields unacceptable fit across multiple fit indices cannot be accepted as a workable

measurement model and any subsequent invariance testing that is conducted will be redundant because the baseline model does not adequately represent the observed variables in the study sample. The stronger the model fit, the more confident the investigator can be that the model provides a valid representation of the latent factors measured by its observed variables. Thus when evaluating a measurement model its conceptual "fit" with prior research, theory and practice should be considered along with its objective fit to the observed variables in the study sample. The measurement model used in these analyses was adopted from Tuokko and colleagues because 1) it has been previously validated in a separate wave of CSHA data and 2) it corresponds well with other established models of cognitive ability.

Although the correspondence of the measurement model with well-validated conceptualizations of the major domains of cognitive ability could be considered a strength, it could also be seen as a weakness. Many neuropsychologists have justifiably moved beyond the conceptualization of independent "domains" of cognitive ability in favour of delineating the lower-level cognitive processes that drive higher-level abilities. For example, processes such as attention and working memory which rely heavily on the integrity of the frontal lobes are essential for performance on tests of memory, visuospatial ability and verbal ability alike. These frontal lobe functions are very sensitive to many forms of brain damage and disease, whereas higher-level abilities such as memory, visuospatial ability and verbal ability are often relatively spared. Thus, in the clinical context where the objective is to characterize the cognitive sequelae of brain insult, measurement models that hold lower-level processes constant by focusing entirely

on higher-level domains do the clinician a disservice by failing to isolate the processes which are most likely to be compromised.

Despite its potential limitations there are a number of justifications for the selected measurement model. First, though the neuropsychological test battery of the CSHA was the most extensive cognitive battery ever to be included in an epidemiologic study protocol, eleven tests remains a relatively small number of observed variables for a CFA model. Following the three indicator per factor rule, three latent factors are the maximum number allowable with eleven indicators. This limitation makes the specification of a more elaborate measurement model impossible. Second, Tuokko and colleagues (2009) tested five different potential measurement models and found the model used in these analyses to demonstrate the best fit to the CSHA data. Finally, it is important to consider that only cognitively intact participants were included in the study sample. In an intact sample the contributions of attention and working memory to performance on tests of higher-level cognitive abilities are of less direct importance because it is assumed (and has been determined by consensus diagnosis) that these processes are intact and functioning normally. The relative contributions of these processes to performance on the eleven observed tests may vary between HE and LE samples, but this issue relates directly to the objective of this thesis because it essentially questions the measurement equivalence of the observed neuropsychological tests.

Model Fit: Configural Invariance

Establishing configural invariance is an important first step in the process of invariance testing because it evaluates whether the measurement model provides a workable characterization in both samples of the patterns of variance observed among the study variables. Configural invariance was tested by first fitting the measurement model to the LE and HE samples independently, and subsequently fitting it to an aggregate sample that included both groups. Although the model demonstrated excellent fit to the LE sample, its fit to the HE sample was suboptimal. Estimates from absolute, parsimony and comparative fit indices all indicated that the fit of the model to the HE sample was barely within acceptable limits. A potential explanation for the discrepancy in model fit between the LE and HE samples is the between-group difference in range of reported years of education. In the LE sample, participants reported between 0 and 8 years of formal education. The HE sample reported a range of 9 to 25 years of education and thus collapses across participants with average-and above-average levels of education. As will be discussed, this may have increased the interindividual variability in the HE sample relative to the LE sample. Though model fit clearly differed between the HE and LE samples, estimates for the aggregate sample indicated good overall fit. Fit estimates were acceptable in both samples as well as in the aggregate sample so the model was accepted as configurally invariant.

As discussed in the Introduction, Paolo and Ryan (1994) determined that a two-factor model best fit the subtests of the WAIS-R in older adults with eleven or fewer years of education while a three-factor model provided a better conceptualization of subtest scores for participants with twelve or more years of education. This finding could

be interpreted as evidence that configural invariance of models of cognitive ability cannot be established across HE and LE samples, which would contradict the findings of the present study. However, the study designed by Paolo and Ryan differs from the present study on an important dimension: In their study, participants with eleven or fewer years of education were classified as LE and those with twelve or more years of education were classified as HE. Thus, Paolo & Ryan compared the performance of well-educated older adults to those with less education by collapsing across those with average and below-average levels of educational attainment. The present study instead compared the performance of adults with little or no formal schooling with performance of those with average and above-average levels of education. This is an important distinction because the effects of education on the measurement equivalence of neuropsychological tests are likely to differ in samples with very low vs. very high levels of education.

The fit of the measurement model to our aggregate CSHA-2 study sample was better than was observed by Tuokko and colleagues (2009) to data from CSHA-1. In the CSHA-1 sample the model did not initially demonstrate acceptable parsimony or fit and post-hoc modifications were required in order to bring these estimates. This contrasts with the present study, where all fit indices were within acceptable limits. One potential explanation for this difference is that this investigation only included cognitively intact participants in the study sample, while Tuokko and colleagues (2009) did not restrict study inclusion based on cognitive status. As was discussed in the Methods, cognitive status compromises the constructs that are measured by neuropsychological tests, and thus a measurement model may be expected to demonstrate poorer fit to a sample that includes participants with cognitive impairment relative to a cognitively intact sample.

Metric Invariance

Metric invariance testing evaluates the uniformity of the association between observed variables and latent factors across groups by placing equality constraints on model factor loadings (Brown, 2006). It is an important component of measurement equivalence because it determines whether group differences in observed test scores can be used to infer the presence of group differences at the latent factor level. Operationally, factor loadings are the regression equations that relate the observed tests to their latent factors and represent the predicted change in an observed test score per unit change in the latent factor (Vandenberg & Lance, 2000). If the influence of a latent factor on any of its associated test scores differs significantly between groups then these test scores are non-invariant. In cases where tests are non-invariant it is not possible to attribute group differences in test performance to differences at the latent factor level.

A stepwise approach to metric invariance testing was used in this thesis, where equality constraints are added incrementally to an initially unconstrained model. Some methodologists have proposed alternative "step-down" approaches to invariance testing where a model with all parameters constrained is fit across groups and parameters are systematically released until the model is found to be invariant (Brown, 2006). However, this approach is less informative and can make it difficult to identify sources of non-invariance when the fully constrained model is ill-fitting. The stepwise approach was thus chosen over a step-down approach because it provides the most prudent method for invariance testing and also facilitates examination of partial invariance (Brown, 2006).

In this study it was not possible to establish full invariance of the factor loadings of the measurement model across HE and LE samples, and thus partial measurement

equivalence was examined by testing the invariance of loadings separately for each of the three factors in the model. Partial invariance analyses revealed that tests measuring Visuospatial Ability and Long-term Retention were invariant across HE and LE samples. In contrast, the factor loadings for Verbal Ability could not be constrained equal across groups. Animal Fluency and the Token test were subsequently identified as the two tests of Verbal Ability that were driving the non-invariance of factor loadings. In the HE sample 43% of variability in Animal fluency scores was accounted for by the Verbal Ability, whereas in the LE sample Verbal Ability accounted for only 26% of the variability in scores. In an opposite trend 13% of variability in performance on the Token Test in the HE sample was accounted for by Verbal Ability and 28% of variability in scores was accounted for by Verbal Ability in the LE sample. Group differences in the association between Verbal Ability and scores on the Token Test and Animal Fluency were statistically significant, thereby rendering these test non-invariant. This means that conclusions regarding verbal ability cannot be drawn based on Animal Fluency or Token Test scores if examinees differ in their levels of education.

None of the eleven observed tests yielded identical loadings across samples. However, only factor loadings for Animal Fluency and the Token Test differed with a magnitude and consistency that was sufficient to reach statistical significance. When constraints on these tests were released, factor loadings for the nine remaining tests could be constrained equal across HE and LE samples. The subsequent addition of equality constraints on factor covariances did not compromise the invariance of the partial invariance model. Together the results of these analyses mean that the nine invariant neuropsychological tests provide an equivalent metric for the three latent factors across

the HE and LE samples, and that the associations among the latent factors are of a similar magnitude in both samples. For example, if mean differences in performance on the Benton Visual Retention test were observed between groups of HE vs. LE older adults, it would be possible to attribute findings to group differences in visuospatial ability, knowing that the test provides a similar metric of visuospatial ability across HE and LE samples and knowing that the integrity of Visuospatial Ability as a standalone psychological construct is comparable across groups.

Some methodologists criticize partial invariance testing: using configural invariance as an example, it is argued that rejection of the null hypothesis of metric invariance should preclude any subsequent investigation of measurement equivalence (Vandenberg & Lance, 2000). It is also argued that because partial invariance testing is always post-hoc it is data-driven and relies heavily on chance. Vandenberg & Lance (2000) reviewed the arguments of critics and proponents of partial invariance testing and recommended that metric invariance constraints only be relaxed a) for a minority of observed variables in a model, b) on a strong theoretical basis and c) when cross-validation evidence points to their viability. They also suggest that non-invariant indicators be removed from analysis.

Invariance constraints in this study were removed from factor loadings for two of eleven observed neuropsychological tests. Although partial invariance testing was not conducted on any a priori theoretical basis, it is defensible for the purposes of this thesis. The model used here serves as a conceptual representation of the cognitive abilities that are measured by the eleven tests in the CSHA neuropsychological battery. The model is not intended for applied use and the values of its latent factors have not been validated as

interpretable indices of cognitive ability. In many contexts, invariance testing is used as a method for validating clinical scales across diverse populations in order to determine whether these scales retain their validity across populations. In these cases, examining partial measurement equivalence without deleting the non-invariant items altogether can make the scoring and interpretation of "partially invariant" scales confusing and potentially meaningless. Although pragmatic as well as conceptual conclusions can be drawn from the results of this study, testing the invariance of factor loadings was conducted as a way of experimentally identifying neuropsychological tests that provide a different metric for their assigned latent variable in the LE vs. HE samples. Subsequent partial invariance testing was conducted as a necessary extension of invariance testing in order to pinpoint non-invariant tests. Cross-validation of findings was not possible due to sample size limitations and thus replication is warranted.

Animal Fluency as a Non-Invariant Measure of Verbal Ability

Animal Fluency is a classic test of semantic fluency that requires examinees to spontaneously produce words under the restriction that all words be the names of animals. This is a multifaceted test that depends on the speed of the examinee (responses must be made within sixty seconds), their ability to monitor the words that they generate and inhibit incorrect responses, their ability to access semantic stores and the size of their semantic stores (e.g., the number of animal names that they know) (Birn et al., 2010). Phonemic and Animal Fluency are similar tasks and place demands on many of the same lower-level cognitive abilities, including sustained attention, inhibition and perceptual speed (Strauss et al., 2006). Where the two tasks differ is in the controlled search strategies that they require: Animal Fluency places demands on semantic knowledge

stores while Phonemic Fluency forces examinees to suppress the habit of using words based on their meaning in favour of using words based on the arbitrary characteristic of what letter that they begin with (Birn et al., 2010).

Both Animal and Phonemic Fluency elicited significant differences in mean group performance. On average, HE participants generated nine more words than LE participants on the Phonemic Fluency test (a 23% advantage) and three more words than LE participants for Animal Fluency (a 12% advantage). The bivariate association between phonemic and animal fluency scores differed between groups, with a Pearson correlation of 0.49 in the HE sample and 0.37 in the LE sample (See Appendix C). This association reveals that Animal and Phonemic Fluency measure more distinctive abilities in LE samples relative to HE samples where there is almost 50% overlap in the psychological constructs and cognitive abilities that are measured by the two tests.

The relative dependence of Animal Fluency on semantic knowledge is reflected in fMRI investigations that demonstrate increased temporal lobe and posterior frontal lobe activation during performance on Animal Fluency relative to Phonemic Fluency, which is predominantly associated with activation of inferior regions of the left frontal lobe (Birn et al., 2010). These findings are also consistent with the known clinical sensitivity of Animal Fluency to early temporal lobe neuropathology associated with disorders such as Alzheimer's disease (Tierney et al., 2005). In these cases Phonemic Fluency scores, which primarily measure frontal lobe function, are well within normal limits despite impaired Animal Fluency. The amount of interindividual variability in the relative contributions of frontal and temporal lobe functions to performance on Animal and Phonemic fluency is not known.

Although Phonemic Fluency scores were invariant across the HE and LE samples, Animal Fluency scores were not. Only 26% of variability in Animal Fluency scores was attributable to verbal ability in the LE sample. In contrast, individual differences in verbal ability account for almost half (43%) of the variability in performance on Animal Fluency in the HE sample. The results of this thesis suggest that variability in Animal Fluency performance in LE samples is due mainly to individual differences in relevant semantic knowledge stores and other non-verbal processes. The stronger association between Animal Fluency and Verbal Ability in HE participants could reflect less variability in the semantic (temporal lobe) demands of Animal Fluency in this sample.

The Token Test as a Non-Invariant Measure of Verbal Ability

The Token Test was developed by Arthur Benton within the context of a comprehensive aphasia battery. Thus, it was explicitly designed to test verbal ability. It provides a measure of receptive language and verbal comprehension by requiring examinees to follow increasingly complex verbal commands regarding the assembly and configuration of plastic tokens (Benton & Hamsher, 1989). The Token Test is sensitive to brain injury and dementia (Millis et al., 2001), but its precise neural correlates have not been explored. The Token Test is a simple task for cognitively intact participants, but a significant difference was observed between HE and LE groups, with HE participants outperforming LE participants by a mean of four points, representing a 10% advantage.

Like many neuropsychological tests, the Token Test requires many cognitive processes other than the receptive language/verbal comprehension abilities that is was designed to measure. Strauss and colleagues (2006) identified short-term memory, inhibitory control and constructional praxis as some of the non-verbal abilities that are

required for performance. The results of this study reveal that the relative contributions of verbal and non-verbal abilities to performance on the Token Test differ significantly across HE and LE samples. As was discussed earlier, the Token Test corresponded more strongly with Verbal Ability in the LE sample relative to the HE sample, where only 13% of variability in test scores were accounted for by Verbal Ability. The verbal aspects of the Token Test are simple for individuals with intact language abilities, and thus it is not surprising that the contribution of Verbal Ability to performance was relatively small even in the LE sample, where 28% of variability in performance was accounted for by individual differences in Verbal Ability. The weaker association between Verbal Ability and Token Test performance in the HE sample relative to the LE sample may be due to the presence of less variability in the verbal aspects of the task in the HE sample.

Between-Group Age and Sex Differences

The LE sample from the CSHA-2 cohort was 2 years older and had a 20% higher proportion of male participants relative to the HE sample. In order to rule out the possibility of artefactual findings due to these demographic differences, all analyses were replicated in a subsample that matched the HE and LE groups on age and a subsample that matched the HE and LE groups on sex. This step was determined to be an important precaution prior to attributing study findings to differences in educational attainment between samples because both age and sex have an impact on cognition.

Evidence of sex differences in cognitive functioning to warranted replication of study findings on a sex-matched sample. Women tend to outperform men on tests of verbal ability and men demonstrate an advantage on tests of visuospatial ability (Weiss et al., 2003). Men also tend to possess higher working memory capacities for object and

spatial information (Lejbak et al., 2011). However, the magnitude of these effects are small and sex differences in neuropsychological test performance are rarely of a clinically significant magnitude (Lezak, 2004). More salient for the present study is evidence of sex differences in strategy implementation during neuropsychological test performance. For instance, Lanting and colleagues (2009) found no sex differences in Animal or Phonemic Fluency performance but noted significant sex differences in the strategies that were adopted for completion of both tasks. Specifically, male participants yielded larger word clusters on both tasks and female participants performed more cluster switches on the Animal Fluency. Thus, despite the absence of group differences in performance on these two tests, there were sex differences in the way the tasks were approached (Lanting et al., 2009). Given the objectives of this thesis it was important to establish that findings were not an artefact of the significantly different sex distributions in the HE and LE samples.

Age is known to negatively impact performance on neuropsychological tests in adult populations, and the magnitude of the association age and cognitive performance is greatest in older adults (Lezak, 2004). The neural correlates of cognitive performance also change as a function of age, with the contributions of parietal lobe functions decreasing and frontal lobe activation increasing (Grady et al., 1994). It has been demonstrated that the increased frontal activation is a compensatory response to decreases in posterior activation that allows for successful cognitive performance (Davis et al., 2008). Age and education are negatively correlated (Lezak, 2004), and because both impact cognitive function it was considered important a priori to examine the age demographics of the HE and LE samples and to implement matched sampling in the event that the groups differed in age.

Results from the age-and sex-matched subsamples suggest that study findings were not an artefact of group differences in age or sex. Estimates of model fit were very similar across all samples, and the same tests emerged as non-invariant in the matched samples as were identified in the full sample. One exception was the Token Test which was invariant in the age-matched sample but non-invariant in the full and sex-matched samples. Nonetheless, as with the other samples, model factor loadings were invariant after removing the equality constraints for Token Test and Animal Fluency. Ideally a single age-and sex-matched sample would have been used to replicate findings, but this was not possible due to sample size limitations. For the purposes of this investigation, replication of findings in the matched samples was considered an acceptable demonstration that results could be attributed to group differences in education.

Potential Limitations

The purpose of this thesis was addressed using the best available participant sample, neuropsychological tests and measurement model. However, each of these critical dimensions had some associated limitations. While none are considered fatal to the integrity of study findings, results should be interpreted within the context of these limitations and future investigations should pursue replication of the observed findings.

The CSHA provided access to a very large representative sample of older adults, but sample size remains a limitation of this study. Specifically, only 120 of the 1,466 English-speaking participants who received a neuropsychological assessment were both cognitively intact and reported less than nine years of formal schooling. Sample size can affect the statistical power and precision of model parameter estimates and a sample size

of 100 is often considered the lower limit for CFA (Brown, 2006). Thus the full LE sample and the matched subsamples were barely within acceptable sample size limits.

The neuropsychological test battery that was used in the CSHA is the most extensive battery ever included in an epidemiologic study protocol. However, many of the neuropsychological tests that were included were abbreviated or slightly modified versions of the tests that are used clinically. For example, abbreviated forms of the Comprehension, Similarities and Block Design WAIS-R subtests were used, and the 20-minute delay period between the immediate and delayed trials of the RAVLT was shortened to 10 minutes. Given the high correlations that are reported between full and abbreviated WAIS subtests in older adults (Dinning & Kraft, 1983), it is unlikely that these changes seriously compromise the generalizability of study findings.

As was discussed earlier, the fit of the model that was selected to represent the psychological constructs measured by the CSHA was acceptable in both samples. However its fit was stronger in the LE sample relative to the HE sample. For the purposes of this thesis the model was accepted as a workable representation of the 11 observed tests in both samples but there were clear between-group differences in model fit. Parameter estimates of the Long-Term Retention factor in the LE sample were also problematic. Specifically, the factor loading for WMS Information was below salient levels. In addition, the extreme kurtosis value for the Token Test was problematic because kurtosis can inflate chi-square values and underestimate CFI and model standard errors. Finally, skewness and kurtosis estimates were to establish univariate normality instead of their critical values. To offset these potential limitations, confidence intervals

and standard error estimates for model parameters were obtained via bootstrapping in order to confirm their stability.

Conclusions

The broad impact of education on cognitive abilities later in life has been recognized ever since Katzman first noted interindividual variability in the association between amyloid deposition and cognitive outcomes in pathologically verified Alzheimer disease (Katzman et al., 1988). Education and other proxies for cognitive reserve have also been studied thoroughly within experimental paradigms. However, until now, the impact of education on the measurement properties of standardized neuropsychological tests was not known.

Until now, there has been no research to support or refute the notion that standardized neuropsychological tests measure the same abilities in older adults regardless of their level of education. This is problematic in the clinical context because it compromises the interpretability of test results of individuals with lower levels of education. The lower-educated demographic is known to be at increased risk of cognitive impairment and dementia, and it is a subject of debate whether their elevated risk is due to exposure to environmental hazards, a true lower baseline level of cognitive ability, or an artefact of tests that are inherently biased due to their educated standardization samples. Further complicating this issue is the possibility that test items are interpreted and approached differently as a function of the examinee's level of education.

The results of this thesis suggest that neuropsychological tests are largely invariant to individual differences in educational attainment. In other words, one can confidently assert that a test of memory or visuospatial ability is measuring the same

cognitive abilities in LE individuals as in HE individuals. This is an important premise for studies that use neuropsychological tests to quantify the protective effects of education on cognitive ability later in life. However, prior to the current study there was no empirical support for this premise.

An important exception to the overall finding of measurement equivalence was the identification of two non-invariant tests of Verbal Ability. The factor loadings for Animal Fluency and the Token Test could not be constrained equal across groups, which mean that performance on these tests does not provide an equal measure of verbal ability in LE and HE individuals. These findings are of pragmatic and conceptual importance.

Pragmatically this thesis provides evidence that Animal fluency and the Token test should be avoided when designing test batteries for studies that stratify their samples by education, or clinically (in the case of Animal Fluency) when attempting to delineate frontal and temporal lobe contributions to a profile of cognitive impairment in clients with little or no formal schooling. More conceptually, this thesis reports that individual differences in strategy implementation and neural resource allocation are strong enough to compromise the abilities and processes that neuropsychological tests measure.

Characterizing relations between brain and behaviour is the domain of neuropsychology. As such, factors that impact or moderate brain-behaviour relationships are of interest because these factors—when not adjusted for—introduce a source of error into behavioural investigations of brain function. This thesis, in conclusion, provides novel evidence of the measurement equivalence of standardized neuropsychological tests to individual differences in education and identifies two tests that cannot provide a reliable measure of verbal ability across HE and LE examinees.

References

Angel, L., Fay, S., Bouazzaoui, B., Baudouin, A., & Isingrini, M. (2010). Protective role of educational level on episodic memory aging: an event-related potential study. *Brain Cogn, 74,* 312-323.

Anstey, K. J., Luszcz, M. A., & Sanchez, L. (2001). A reevaluation of the common factor theory of shared variance among age, sensory function, and cognitive function in older adults. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences, 56,* 3-11.

Baltes, P. B. & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J.R.Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Introduction to research methodological issues* (New York: Academic Press.

Bartres-Faz, D., Sole-Padulles, C., Junque, C., Rami, L., Bosch, B., Bargallo, N. et al. (2009). Interactions of cognitive reserve with regional brain anatomy and brain function during a working memory task in healthy elders. *Biol.Psychol., 80,* 256-259.

Benson, J. & Fleishman, J. A. (1994). The Robustness of Maximum-Likelihood and Distribution-Free Estimators to Nonnormality in Confirmatory Factor-Analysis. *Quality & Quantity, 28,* 117-136.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol.Bull., 107,* 238-246.

Benton, A. L. (1974). *Revised Visual Retention test: clinical and experimental applications*. (4 ed.) New York: Psychological Corporation.

Benton, A. L. & Hamsher, K. (1989). *Multilingual Aphasia Examination*. (2 ed.) Iowa City: Iowa USA: AJA Associates.

Birn, R. M., Kenworthy, L., Case, L., Caravella, R., Jones, T. B., Bandettini, P. A. et al. (2010). Neural systems supporting lexical search guided by letter and semantic category cues: a self-paced overt response fMRI study of verbal fluency. *Neuroimage., 49,* 1099-1107.

Bleecker, M. L., Bollawilson, K., Agnew, J., & Meyers, D. A. (1988). Age-Related Sex-Differences in Verbal Memory. *Journal of Clinical Psychology, 44,* 403-411.

Bollen, K. A. (1989).

Structural equations with latent variables. New York: Wiley.

Boone, K. B., Ponton, M. O., Gorsuch, R. L., Gonzalez, J. J., & Miller, B. L. (1998). Factor analysis of four measures of prefrontal lobe functioning. *Arch.Clin.Neuropsychol., 13,* 585-595.

Bosch, B., Bartres-Faz, D., Rami, L., renaza-Urquijo, E. M., Fernandez-Espejo, D., Junque, C. et al. (2010). Cognitive reserve modulates task-induced activations and

deactivations in healthy elders, amnestic mild cognitive impairment and mild Alzheimer's disease. *Cortex, 46,* 451-461.

Bowden, S. C., Dodds, B., Whelan, G., Long, C., Dudgeon, P., Ritter, A. et al. (1997). Confirmatory factor analysis of the Wechsler Memory Scale Revised in a sample of clients with alcohol dependency. *Journal of Clinical and Experimental Neuropsychology, 19,* 755-762.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guildford Press.

Brucki, S. M. & Nitrini, R. (2008). Cancellation task in very low educated people. *Arch.Clin.Neuropsychol., 23,* 139-147.

Buckner, R. L., ndrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann.N.Y.Acad.Sci., 1124,* 1-38.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 155-175.

Coman, E., Moses, J. A., Jr., Kraemer, H. C., Friedman, L., Benton, A. L., & Yesavage, J. (2002). Interactive influences on BVRT performance level: geriatric considerations. *Arch.Clin.Neuropsychol., 17,* 595-610.

Crossley, M., D'Arcy, C., & Rawson, N. S. (1997). Letter and category fluency in community-dwelling Canadian seniors: a comparison of normal participants to those with dementia of the Alzheimer or vascular type. *J.Clin.Exp.Neuropsychol., 19,* 52-62.

Crosson, B., Sartor, K. J., & Jenny, A. B. (1990). Mr Localization and Memory in Chronic Head-Injured Adults. *Journal of Clinical and Experimental Neuropsychology, 12,* 34.

Davis, R. N., Massman, P. J., & Doody, R. S. (2003). WAIS-R factor structure in Alzheimer's disease patients: a comparison of alternative models and an assessment of their generalizability. *Psychol.Aging, 18,* 836-843.

Davis, S. W., Dennis, N. A., Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2008). Que PASA? The posterior-anterior shift in aging. *Cereb.Cortex, 18,* 1201-1209.

Dinning, W. D. & Kraft, W. A. (1983). Validation of the Satz-Mogel short form for the WAIS-R with psychiatric inpatients. *J.Consult Clin.Psychol., 51,* 781-782.

Dunlosky, J. & Salthouse, T. A. (1996). A decomposition of age-related differences in multitrial free recall. *Aging Neuropsychology and Cognition, 3,* 2-14.

Elias, M. F., Elias, P. K., D'Agostino, R. B., Silbershatz, H., & Wolf, P. A. (1997). Role of age, education, and gender on cognitive performance in the Framingham Heart Study: community-based norms. *Exp.Aging Res., 23,* 201-235.

Fabrigoule, C., Rouch, I., Taberly, A., Letenneur, L., Commenges, D., Mazaux, J. M. et al. (1998). Cognitive process in preclinical phase of dementia. *Brain, 121 (Pt 1),* 135-141.

Filley, C. M., Heaton, R. K., Nelson, L. M., Burks, J. S., & Franklin, G. M. (1989). A comparison of dementia in Alzheimer's disease and multiple sclerosis. *Arch.Neurol., 46,* 157-161.

Flanagan, D. P. & Harrison, P. L. (2005). *Contemporary intellectual assessment: Theories, tests, and issues*. (2nd ed.) New York: The Guilford Press.

Grady, C. L., Maisog, J. M., Horwitz, B., Ungerleider, L. G., Mentis, M. J., Salerno, J. A. et al. (1994). Age-Related-Changes in Cortical Blood-Flow Activation During Visual Processing of Faces and Location. *Journal of Neuroscience, 14,* 1450-1462.

Hall, K., Gureje, O., Gao, S., Ogunniyi, A., Hui, S. L., Baiyewu, O. et al. (1998). Risk factors and Alzheimer's disease: a comparative study of two communities. *Aust.N.Z.J.Psychiatry, 32,* 698-706.

Heaton, R. K., Matthews, C. G., Grant, I., & Avitable, N. (1996). Demographic corrections with comprehensive norms: An overzealous attempt or a good start? *Journal of Clinical and Experimental Neuropsychology, 18,* 449-458.

Henry, J. D. & Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology, 18,* 284-295.

Henry, J. D. & Crawford, J. R. (2005). A meta-analytic review of verbal fluency deficits in schizophrenia relative to other neurocognitive deficits. *Cogn Neuropsychiatry, 10,* 1-33.

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia, 42,* 1212-1222.

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.

Ivnik, R. J., Malec, J. F., & Smith, G. E. (1992). Mayo's Older Americans Normative Studies: WAIS-R Norms for Ages 56-97. *The Clinical Neuropsychologist, 6,* 1-30.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., & Petersen, R. C. (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. *Clinical Neuropsychologist, 10,* 262-278.

Joreskog, K. G. & Sorbom, D. (1988).

LISREL 7: A guide to the program and applications. Chicago: SPSS.

Kantarci, K., Jack, C. R., Xu, Y. C., Campeau, N. G., O'Brien, P. C., Smith, G. E. et al. (2001). Mild cognitive impairment and Alzheimer disease: Regional diffusivity of water. *Radiology, 219,* 101-107.

Kantarci, K., Petersen, R. C., Boeve, B. F., Knopman, D. S., Weigand, S. D., O'Brien, P. C. et al. (2005). DWI predicts future progression to Alzheimer disease in amnestic mild cognitive impairment. *Neurology, 64,* 902-904.

Karzmark, P., Heaton, R. K., Grant, I., & Matthews, C. G. (1984). Use of demographic variables to predict overall level of performance on the Halstead-Reitan Battery. *J.Consult Clin.Psychol., 52,* 663-665.

Katzman, R., Terry, R., DeTeresa, R., Brown, T., Davies, P., Fuld, P. et al. (1988). Clinical, pathological, and neurochemical changes in dementia: a subgroup with preserved mental status and numerous neocortical plaques. *Ann.Neurol., 23,* 138-144.

Kaufman, A. S., Lichtenberger, E. O., & McLean, J. E. (2001). Two- and three-factor solutions of the WAIS-III. *Assessment., 8,* 267-280.

Kawas, C. H., Corrada, M. M., Brookmeyer, R., Morrison, A., Resnick, S. M., Zonderman, A. B. et al. (2003). Visual memory predicts Alzheimer's disease more than a decade before diagnosis. *Neurology, 60,* 1089-1093.

Lanting, S., Haugrud, N., & Crossley, M. (2009). The effect of age and sex on clustering and switching during speeded verbal fluency tasks. *J.Int.Neuropsychol.Soc., 15,* 196-204.

Lass, N. J., Depaolo, A. M., Simcoe, J. C., & Samuel, S. M. (1975). Normative Study of Childrens Performance on Short Form of Token Test. *Journal of Communication Disorders, 8,* 193-198.

Lazarov, O., Robinson, J., Tang, Y. P., Hairston, I. S., Korade-Mirnics, Z., Lee, V. M. et al. (2005). Environmental enrichment reduces Abeta levels and amyloid deposition in transgenic mice. *Cell, 120,* 701-713.

Le Carret, N., Lafont, S., Mayo, W., & Fabrigoule, C. (2003a). The effect of education on cognitive performances and its implication for the constitution of the cognitive reserve. *Dev.Neuropsychol., 23,* 317-337.

Le Carret, N., Rainville, C., Lechevallier, N., Lafont, S., Letenneur, L., & Fabrigoule, C. (2003b). Influence of education on the benton visual retention test performance as mediated by a strategic search component. *Brain Cogn, 53,* 408-411.

Lejbak, L., Crossley, M., & Vrbancic, M. (2011). A male advantage for spatial and object but not verbal working memory using the n-back task. *Brain Cogn, 76,* 191-196.

Levi, O., Jongen-Relo, A. L., Feldon, J., Roses, A. D., & Michaelson, D. M. (2003). ApoE4 impairs hippocampal plasticity isoform-specifically and blocks the environmental stimulation of synaptogenesis and memory. *Neurobiology of Disease, 13,* 273-282.

Lezak, M. L. (2004). *Neuropsychological Assessment*. (4 ed.) Oxford University Press.

Loehlin, J. C. (1988). *Latent variable models: An introduction to factor, path, and structural analysis*. (3 ed.) Mahwah, N.J.: Lawrence Erlbaum.

Mazaux, J. M., Dartigues, J. F., Letenneur, L., Darriet, D., Wiart, L., Gagnon, M. et al. (1995). Visuospatial Attention and Psychomotor Performance in Elderly

Community Residents - Effects of Age, Gender, and Education. *Journal of Clinical and Experimental Neuropsychology, 17,* 71-81.

McCarthy, F. M., Sellers, A. H., Burns, W. J., Smith, G., Ivnik, R. J., & Malec, J. F. (2003). Prediction of IQ in the Mayo Older Adult Normative sample using multiple methods. *J.Clin.Psychol., 59,* 457-463.

McDowell, I., Hill, G., & Lindsay, J. (2001). An overview of the Canadian Study of Health and Aging. *Int.Psychogeriatr., 13 Supp 1,* 7-18.

McDowell, I., Xi, G., Lindsay, J., & Tierney, M. (2007). Mapping the connections between education and dementia. *J.Clin.Exp.Neuropsychol., 29,* 127-141.

Miatton, M., Wolters, M., Lannoo, E., & Vingerhoets, G. (2004). Updated and extended Flemish normative data of commonly used neuropsychological tests. *Psychologica Belgica, 44,* 189-216.

Millis, S. R., Rosenthal, M., Novack, T. A., Sherer, M., Nick, T. G., Kreutzer, J. S. et al. (2001). Long-term neuropsychological outcome after traumatic brain injury. *Journal of Head Trauma Rehabilitation, 16,* 343-355.

Mirochnic, S., Wolf, S., Staufenbiel, M., & Kempermann, G. (2009). Age effects on the regulation of adult hippocampal neurogenesis by physical activity and environmental enrichment in the APP23 mouse model of Alzheimer disease. *Hippocampus, 19,* 1008-1018.

Mitrushina, M., Satz, P., Chervinsky, A., & Delia, L. (1991). Performance of 4 Age-Groups of Normal Elderly on the Rey Auditory-Verbal Learning Test. *Journal of Clinical Psychology, 47,* 351-357.

Moses, J. A., Jr. (1986). Factor structure of Benton's tests of visual retention, visual construction, and visual form discrimination. *Arch.Clin.Neuropsychol., 1,* 147-156.

Osorio, A., Fay, S., Pouthas, V., & Ballesteros, S. (2010). Ageing affects brain activity in highly educated older adults: An ERP study using a word-stem priming task. *Cortex, 46,* 522-534.

Paolo, A. M. & Ryan, J. J. (1994). Factor structure of the WAIS-R by educational level: an examination of elderly persons. *Arch.Clin.Neuropsychol., 9,* 259-264.

Piras, F., Cherubini, A., Caltagirone, C., & Spalletta, G. (2011). Education mediates microstructural changes in bilateral hippocampus. *Hum.Brain Mapp., 32,* 282-289.

Read, D. E. (1980). *Neuropsychological Assessment of Memory in Early Dementia: Normative data for a new battery of memory tests "Animal Naming"*. Victoria, British Columbia: University of Victoria.

Rodriguez-Aranda, C. & Sundet, K. (2006). The frontal hypothesis of cognitive aging: factor structure and age effects on four frontal tests among healthy individuals. *J.Genet.Psychol., 167,* 269-287.

Rosen, V. M. & Engle, R. W. (1997). The role of working memory capacity in retrieval. *J.Exp.Psychol.Gen., 126,* 211-227.

Satz, P. & Mogel, S. (1962). An abbreviation of the WAIS for clinical use. *J.Clin.Psychol., 18,* 77-79.

Scarmeas, N., Zarahn, E., Anderson, K. E., Hilton, J., Flynn, J., Van Heertum, R. L. et al. (2003). Cognitive reserve modulates functional brain responses during memory tasks: a PET study in healthy young and elderly subjects. *Neuroimage., 19,* 1215-1227.

Sliwinski, M. & Buschke, H. (1999). Cross-sectional and longitudinal relationships among age, cognition, and processing speed. *Psychol.Aging, 14,* 18-33.

Spreen, O. & Benton, A. L. (1977). *Neurosensory center comprehensive examination for aphasia*. Victoria, British Columbia: University of Victoria.

Springer, M. V., McIntosh, A. R., Winocur, G., & Grady, C. L. (2005). The relation between brain activity during memory tasks and years of education in young and older adults. *Neuropsychology., 19,* 181-192.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research, 25,* 173-180.

Stern, Y. (2009). Cognitive reserve. *Neuropsychologia, 47,* 2015-2028.

Stern, Y., Habeck, C., Moeller, J., Scarmeas, N., Anderson, K. E., Hilton, H. J. et al. (2005). Brain networks associated with cognitive reserve in healthy young and old adults. *Cereb.Cortex, 15,* 394-402.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. (3rd ed.) New York: Oxford University Press.

Teng, E. L. & Chui, H. C. (1987). The Modified Mini-Mental State (3MS) examination. *J.Clin.Psychiatry, 48,* 314-318.

Teresi, J. A. (2006). Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Med.Care, 44,* S39-S49.

Tierney, M. C., Yao, C., Kiss, A., & McDowell, I. (2005). Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. *Neurology, 64,* 1853-1859.

Tuokko, H. & Crockett, D. (1989). Cued-Recall and Memory Disorders in Dementia. *Journal of Clinical and Experimental Neuropsychology, 11,* 278-294.

Tuokko, H., Garrett, D. D., McDowell, I., Silverberg, N., & Kristjansson, B. (2003). Cognitive decline in high-functioning older adults: reserve or ascertainment bias? *Aging Ment.Health, 7,* 259-270.

Tuokko, H., Kristjansson, E., & Miller, J. (1995). Neuropsychological Detection of Dementia - An Overview of the Neuropsychological Component of the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology, 17,* 352-373.

Tuokko, H., Vernonwilkinson, R., Weir, J., & Beattie, B. L. (1991). Cued-Recall and Early Identification of Dementia. *Journal of Clinical and Experimental Neuropsychology, 13,* 871-879.

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. *J.Int.Neuropsychol.Soc., 15,* 416-425.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's verbal learning test: Normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11,* 290-302.

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3,* 4-70.

Warrington, E. K., James, M., & Maciejewski, C. (1986). The WAIS as a lateralizing and localizing diagnostic instrument: a study of 656 patients with unilateral cerebral lesions. *Neuropsychologia, 24,* 223-239.

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

Wechsler, D. & Stone, C. P. (1974). *Wechsler Memory Scale Manual*. New York: Psychological Corporation.

Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W., & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and Individual Differences, 35,* 863-875.

Yen, Y. C., Yang, M. J., Shih, C. H., & Lung, F. W. (2004). Cognitive impairment and associated risk factors among aged community members. *Int.J.Geriatr.Psychiatry, 19,* 564-569.

APPENDIX A

Table 5b. Univariate and multivariate characteristics of the 11 observed variables in the age-matched sample (n=115).

Sample	LE Sample				HE Sample			
Variable	Min	Max	Skew	Kurt	Min	Max	Skew	Kurt
Information	1	6	-1.00	0.06	1	6	-1.32	1.21
Buschke Recall	13	35	-0.16	-0.11	11	35	-0.74	1.23
Rey AVLT	13	63	0.65	0.62	19	67	0.41	-0.18
Digit Symbol	5	60	0.67	1.63	9	62	0.35	-0.08
Block Design	0	20	0.24	0.05	0	24	0.55	1.54
BVRT	2	16	-0.64	0.14	7	16	-0.26	-0.37
Phonemic Fluency	3	55	0.81	0.48	9	60	0.27	-0.51
Comprehension	2	15	-0.17	-0.51	2	16	-0.42	0.11
Token Test	16	66	0.86	3.87	24	66	1.21	5.55
Animal Fluency	6	25	0.68	0.28	4	27	0.28	-0.14
Similarities	0	14	0.28	-1.13	0	14	-0.43	-0.93
Mardia				7.03				3.682

Note. LE = lower-educated sample; HE = higher-educated sample; WMS = Weschler Memory Scales; AVLT = Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales.

Table 5c. Univariate and multivariate characteristics of the 11 observed variables in the sex-matched sample (n=120).

Sample	LE Sample				HE Sample			
Variable	Min	Max	Skew	Kurt	Min	Max	Skew	Kurt
Information	1	6	-1.00	0.06	1	6	-1.32	1.21
Buschke Recall	13	35	-0.16	-0.11	11	35	-0.74	1.23
Rey AVLT	13	63	0.65	0.62	19	67	0.41	-0.18
Digit Symbol	5	60	0.67	1.63	9	62	0.35	-0.08
Block Design	0	20	0.24	0.05	0	24	0.55	1.54
BVRT	2	16	-0.64	0.14	7	16	-0.26	-0.37
Phonemic Fluency	3	55	0.81	0.48	9	60	0.27	-0.51
Comprehension	2	15	-0.17	-0.51	2	16	-0.42	0.11
Token Test	16	66	0.86	3.87	24	66	1.21	5.55
Animal Fluency	6	25	0.68	0.28	4	27	0.28	-0.14
Similarities	0	14	0.28	-1.13	0	14	-0.43	-0.93
Mardia				7.03				3.682

Note. LE = lower-educated sample; HE = higher-educated sample; WMS = Weschler Memory Scales; AVLT = Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales.

APPENDIX B

Bivariate associations among the 11 observed variables in the full study sample (n=505).

	1	2	3	4	5	6	7	8	9	10	11
1. Token Test	1	0.32	0.35	0.41	0.38	0.23	0.16	0.11	0.28	0.08	0.21
2. Phonemic Fluency	0.55	1	0.49	0.37	0.39	0.48	0.24	0.16	0.32	0.23	0.43
3. Animal Fluency	0.32	0.37	1	0.37	0.39	0.49	0.31	0.19	0.32	0.26	0.38
4. WAIS Comprehension	0.46	0.46	0.32	1	0.59	0.34	0.31	0.16	0.26	0.15	0.33
5. WAIS Similarities	0.49	0.61	0.35	0.39	1	0.42	0.29	0.15	0.35	0.18	0.34
6. WAIS Digit Symbol	0.47	0.56	0.46	0.33	0.56	1	0.46	0.32	0.39	0.34	0.41
7. WAIS Block Design	0.42	0.41	0.27	0.26	0.38	0.47	1	0.16	0.40	0.21	0.21
8. WMS Information	0.04	0.12	0.17	0.00	0.12	0.28	0.11	1	0.22	0.26	0.26
9. Benton Visual Retention	0.38	0.41	0.26	0.33	0.43	0.45	0.32	0.02	1	0.24	0.28
10. Buschke Cued Recall	0.11	0.25	0.29	0.11	0.09	0.34	0.24	0.26	0.17	1	0.49
11. Rey AVLT	0.33	0.46	0.35	0.22	0.41	0.52	0.37	0.11	0.39	0.58	1

HE Sample Correlations

LE Sample Correlations

Note. LE = lower-educated sample; HE = higher-educated sample; Rey AVLT = Rey Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales.

Bivariate associations among the 11 observed variables in the age-matched sample (n=230).

	1	2	3	4	5	6	7	8	9	10	11
1. Token Test	1	0.14	0.30	0.33	0.28	0.21	0.18	0.01	0.22	0.09	0.12
2. Phonemic Fluency	0.39	1	0.43	0.29	0.39	0.42	0.28	0.09	0.25	0.25	0.40
3. Animal Fluency	0.25	0.38	1	0.37	0.38	0.46	0.36	0.14	0.34	0.31	0.39
4. WAIS Comprehension	0.41	0.49	0.29	1	0.54	0.21	0.23	0.15	0.15	0.13	0.32
5. WAIS Similarities	0.45	0.62	0.32	0.39	1	0.43	0.22	0.12	0.30	0.20	0.20
6. WAIS Digit Symbol	0.43	0.56	0.45	0.33	0.55	1	0.47	0.32	0.44	0.47	0.37
7. WAIS Block Design	0.38	0.39	0.25	0.27	0.36	0.46	1	0.20	0.52	0.32	0.33
8. WMS Information	0.07	0.11	0.14	0.01	0.10	0.27	0.09	1	0.27	0.40	0.22
9. Benton Visual Retention	0.33	0.41	0.24	0.32	0.42	0.45	0.31	0.00	1	0.42	0.37
10. Buschke Cued Recall	0.11	0.22	0.28	0.13	0.06	0.32	0.21	0.24	0.17	1	0.44
11. Rey AVLT	0.34	0.45	0.31	0.32	0.38	0.50	0.35	0.09	0.38	0.58	1
	LE Sample Correlations										

HE Sample Correlations

Note. LE = lower-educated sample; HE = higher-educated sample; Rey AVLT = Rey Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales.

Bivariate associations among the 11 observed variables in the sex-matched sample (n=240).

	1	2	3	4	5	6	7	8	9	10	11
1. Token Test	1	0.08	0.29	0.31	0.26	0.21	0.07	0.08	0.20	0.11	0.17
2. Phonemic Fluency	0.37	1	0.50	0.43	0.41	0.52	0.18	0.23	0.26	0.26	0.41
3. Animal Fluency	0.20	0.37	1	0.38	0.52	0.61	0.20	0.18	0.35	0.29	0.41
4. WAIS Comprehension	0.36	0.47	0.32	1	0.67	0.42	0.26	0.19	0.24	0.11	0.23
5. WAIS Similarities	0.40	0.60	0.35	0.39	1	0.51	0.23	0.15	0.42	0.14	0.32
6. WAIS Digit Symbol	0.39	0.56	0.46	0.33	0.56	1	0.37	0.30	0.43	0.37	0.41
7. WAIS Block Design	0.33	0.39	0.27	0.26	0.38	0.47	1	0.15	0.42	0.26	0.28
8. WMS Information	0.06	0.12	0.16	0.00	0.12	0.28	0.11	1	0.27	0.31	0.17
9. Benton Visual Retention	0.30	0.40	0.25	0.33	0.42	0.45	0.31	0.01	1	0.27	0.30
10. Buschke Cued Recall	0.10	0.25	0.29	0.11	0.09	0.34	0.24	0.26	0.16	1	0.56
11. Rey AVLT	0.30	0.46	0.35	0.34	0.42	0.52	0.37	0.11	0.39	0.58	1

HE Sample Correlations

LE Sample Correlations

Note. LE = lower-educated sample; HE = higher-educated sample; Rey AVLT = Rey Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales.

APPENDIX C

Table 6b. Configural invariance: Baseline model fit in age-matched groups

Sample	X^2	*df*	Pars.	CFI	RMSEA (90% CI)
Aggregate sample (n=230)	108.45	82	1.32	0.96	0.04 (0.01-0.06)
HE sample only (n=115)	63.17	41	1.54	0.93	0.07 (0.03-0.10)
LE sample only (n=115)	45.28	41	1.1	0.99	0.03 (0.00-0.07)

Note. LE = low-educated; HE = high-educated; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval.

Table 6c. Configural invariance: Baseline model fit in sex-matched groups.

Sample	X^2	*df*	Pars.	CFI	RMSEA (90% CI)
Aggregate sample (n=240)	124.25	82	1.52	0.94	0.05 (0.03-0.06)
HE sample only (n=120)	80.26	41	1.96	0.90	0.09 (0.06-0.12)
LE sample only (n=120)	44.02	41	1.07	0.99	0.03 (0.00-0.07)

Note. LE = low-educated; HE = high-educated; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval.

APPENDIX D

Table 7b. Squared multiple correlations, standardized and unstandardized coefficients for the model when applied to the age-matched HE and LE samples.

Construct	Variable	Sample	SMC (95% CI)	B (95% CI)	β (95% CI)	SE
Verbal Ability	Phonemic Fluency	HE	0.34 (0.16-0.52)	3.38 (1.99-5.55)	0.58 (0.40-0.72)	0.72
		LE	0.62 (0.46-0.76)	5.21 (3.94-7.99)	0.79 (0.66-0.87)	0.89
	Animal Fluency	HE	0.45 (0.23-0.64)	0.48 (0.32-0.69)	0.67 (0.48-0.80)	0.10
		LE	0.24 (0.08-0.38)	0.22 (0.13-0.31)	0.49 (0.28-0.61)	0.04
	Token Test	HE	0.17 (0.02-0.39)	0.43 (0.14-0.81)	0.41 (0.15-0.63)	0.12
		LE	0.34 (0.16-0.52)	0.51 (0.36-0.77)	0.58 (0.39-0.72)	0.09
	Comprehension	HE	0.36 (0.17-0.52)	0.30 (0.19-0.48)	0.60 (0.42-0.72)	0.06
		LE	0.32 (0.17-0.52)	0.19 (0.13-0.27)	0.57 (0.41-0.72)	0.03
	Similarities	HE	0.46 (0.26-0.68)	0.41 (0.26-0.70)	0.68 (0.51-0.82)	0.08
		LE	0.56 (0.41-0.72)	0.33 (0.27-0.42)	0.75 (0.64-0.85)	0.04
Visuospa-tial Ability	Digit Symbol	HE	0.56 (0.35-0.76)	2.79 (1.73-4.16)	0.75 (0.59-0.87)	0.47
		LE	0.62 (0.46-0.76)	2.91 (2.25-4.27)	0.79 (0.68-0.87)	0.51
	Block Design	HE	0.40 (0.21-0.57)	0.36 (0.24-0.53)	0.64 (0.46-0.76)	0.06
		LE	0.32 (0.18-0.47)	0.34 (0.25-0.45)	0.56 (0.43-0.68)	0.06
	BVRT	HE	0.43 (0.22-0.61)	0.18 (0.12-0.23)	0.66 (0.47-0.78)	0.03
		LE	0.34 (0.21-0.49)	0.22 (0.16-0.31)	0.58 (0.46-0.70)	0.04
Long-term Retention	Buschke	HE	0.54 (0.26-0.77)	0.53 (0.28-0.88)	0.73 (0.51-0.88)	0.16
		LE	0.33 (0.00-0.46)	0.29 (0.04-0.46)	0.58 (0.00-0.68)	0.10
	Information	HE	0.22 (0.05-0.45)	0.17 (0.08-0.24)	0.47 (0.22-0.67)	0.41
		LE	0.01 (0.00-0.04)	0.04 (0.01-0.14)	0.09 (0.00-0.19)	0.05
	Rey AVLT	HE	0.37 (0.16-0.63)	1.88 (1.16-3.61)	0.61 (0.41-0.80)	0.55
		LE	0.99 (0.81-1.00)	3.46 (2.21-15.2)	0.99 (0.90-1.00)	2.53

Note. LE = lower-educated; HE = higher-educated; Rey AVLT = Rey Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales; SMC = squared multiple correlations; β = standardized regression weights; B = unstandardized regression weights; CI = confidence intervals.

Table 7c. Squared multiple correlations, standardized and unstandardized coefficients for the model when applied to the sex-matched HE and LE samples.

Construct	Variable	Sample	SMC (95% CI)	B (95% CI)	β (95% CI)	SE
Verbal Ability	Phonemic Fluency	HE	0.34 (0.14-0.49)	3.38 (1.99-5.55)	0.58 (0.37-0.70)	0.90
		LE	0.62 (0.43-0.74)	5.21 (0.89-5.84)	0.79 (0.66-0.86)	0.05
	Animal Fluency	HE	0.51 (0.34-0.66)	0.42 (0.34-0.57)	0.71 (0.59-0.81)	0.06
		LE	0.26 (0.13-0.40)	0.24 (0.15-0.36)	0.51 (0.36-0.63)	0.05
	Token Test	HE	0.11 (0.02-0.24)	0.30 (0.13-0.48)	0.33 (0.14-0.49)	0.09
		LE	0.27 (0.11-0.46)	0.49 (0.36-0.79)	0.52 (0.34-0.68)	0.10
	Comprehension	HE	0.49 (0.27-0.66)	0.28 (0.19-0.42)	0.70 (0.52-0.81)	0.05
		LE	0.31 (0.13-0.50)	0.19 (0.11-0.28)	0.56 (0.36-0.71)	0.04
	Similarities	HE	0.62 (0.40-0.76)	0.38 (0.29-0.58)	0.79 (0.64-0.87)	0.07
		LE	0.56 (0.39-0.74)	0.34 (0.28-0.43)	0.75 (0.63-0.86)	0.04
Visuospa-tial Ability	Digit Symbol	HE	0.56 (0.34-0.76)	0.27 (0.47-0.59)	0.75 (0.59-0.87)	0.60
		LE	0.62 (0.47-0.78)	2.91 (0.51-5.70)	0.79 (0.68-0.88)	0.05
	Block Design	HE	0.22 (0.09-0.37)	0.26 (0.16-0.38)	0.47 (0.30-0.61)	0.06
		LE	0.32 (0.20-0.48)	0.34 (0.25-0.46)	0.57 (0.44-0.70)	0.06
	BVRT	HE	0.36 (0.19-0.56)	0.13 (0.09-0.19)	0.60 (0.43-0.75)	0.02
		LE	0.33 (0.19-0.46)	0.21 (0.15-0.29)	0.58 (0.44-0.68)	0.04
Long-term Retention	Buschke	HE	0.52 (0.27-0.88)	0.46 (0.24-0.83)	0.72 (0.52-0.94)	0.17
		LE	0.35 (0.10-0.53)	0.31 (0.10-0.46)	0.59 (0.30-0.73)	0.09
	Information	HE	0.13 (0.01-0.39)	0.05 (0.10-0.12)	0.36 (0.08-0.62)	0.03
		LE	0.02 (0.00-0.13)	0.02 (0.01-0.07)	0.13 (0.01-0.36)	0.02
	Rey AVLT	HE	0.37 (0.15-0.61)	1.87 (0.37-0.51)	0.61 (0.39-0.78)	0.55
		LE	0.99 (0.64-3.72)	3.46 (0.80-4.33)	0.99 (0.80-1.93)	0.24

Note. LE = lower-educated; HE = higher-educated; Rey AVLT = Rey Auditory Verbal Learning Test; WAIS = Weschler Adult Intelligence Scales; WMS = Weschler Memory Scales; SMC = squared multiple correlations; β = standardized regression weights; B = unstandardized regression weights; CI = confidence intervals.

APPENDIX E

Table 9b. Metric invariance of the measurement model across age-matched LE and HE samples.

Model	X^2	*df*	*p*-value	Invariant?
Metric invariance				
A. All factors constrained	20.67	8	0.01	No
B. Long-term Retention released	12.99	6	0.04	No
C. Visuospatial Ability released	19.76	6	0.003	No
D. Verbal Ability released	8.75	4	0.07	**Yes**
Partial metric invariance: Verbal Ability				
E. Similarities released	20.67	7	0.004	No
F. Comprehension released	19.28	7	0.01	No
G. Phonemic fluency released	17.21	7	0.02	No
H. Animal fluency released	12.82	7	0.08	**Yes**
I. Token test released	17.35	7	0.02	No
Partial invariance: All factors constrained*				
J. Factor loadings only	12.82	6	0.08	**Yes**
K. Factor loadings + Structural covariances	17.44	12	0.13	**Yes**

Note. LE = lower-educated; HE = higher-educated; df = degrees of freedom; * non-invariant tests released.

Table 9c. Metric invariance of the measurement model across sex-matched LE and HE samples.

Model	X^2	*df*	*p*-value	Invariant?
Metric invariance				
A. All factors constrained	19.29	8	0.01	No
B. Long-term Retention released	16.79	6	0.01	No
C. Visuospatial Ability released	15.76	6	0.02	No
D. Verbal Ability released	6.54	4	0.16	**Yes**
Partial metric invariance: Verbal Ability				
E. Similarities released	18.97	7	0.001	No
F. Comprehension released	16.69	7	0.02	No
G. Phonemic fluency released	16.79	7	0.01	No
H. Animal fluency released	13.70	7	0.06	**Yes**
I. Token test released	13.97	7	0.054	**Yes**
Partial invariance: All factors constrained*				
J. Factor loadings only	9.63	6	0.14	**Yes**
K. Factor loadings + Structural covariances	10.82	12	0.55	**Yes**

Note. LE = lower-educated; HE = higher-educated; df = degrees of freedom; * non-invariant tests released.