American Psychological Association

Neuropsychology

ISSN: 0894-4105

2021, Vol. 35, No. 4, 434–450
https://doi.org/10.1037/neu0000736

Factor Analyses of the ADNI Neuropsychological Battery: An Examination of Diagnostic and Longitudinal Invariance

M. Chang and C. J. Brainerd
Department of Psychology and Human Neuroscience Institute, Cornell University

Objective: In the Alzheimer's Disease Neuroimaging Initiative (ADNI), cognitive function was tracked across multiple years by a comprehensive neuropsychological battery. In this study, we examined the latent structure of the ADNI battery and evaluated the invariance of that structure among diagnostic groups and over time. ***Method:*** We used exploratory and confirmatory factor analyses to investigate the invariance of the ADNI battery's latent factor structure among three diagnostic groups (healthy controls, patients with mild cognitive impairment, patients with Alzheimer's disease) over a 2-year interval (baseline, 6 months, 12 months, 24 months). ***Results:*** The results revealed a five-factor structure for the ADNI battery (memory, visuospatial processing, attention, language, executive function). This structure displayed configural invariance but not weak, strong, or strict invariance across the three diagnostic groups. Longitudinally, configural, weak, strong, and strict invariance were all established within each diagnostic group, except that strict invariance was rejected in healthy controls. ***Conclusions:*** The ADNI battery assesses the same cognitive abilities in the three diagnostic groups, but test scores do not calibrate to these abilities equally in the respective groups, making certain statistics (e.g., factor scores) noncomparable between groups. Within each group, the latent structure and the numerical relations between individual tests and underlying factors remained invariant over 2 years, suggesting that this battery is a reliable tool for tracking longitudinal changes in specific cognitive abilities within individual diagnostic groups.

Key Points

Question: Does the ADNI neuropsychological battery measure the same cognitive functions in the same way across diagnostic groups and over time? ***Findings:*** The ADNI battery measures the same cognitive functions in the same way over time, but in different ways across diagnostic groups. ***Importance:*** It is inappropriate to directly compare composite scores or factor scores of the ADNI battery across diagnostic groups. ***Next Steps:*** Future studies are recommended to examine the validity of between-group comparisons with other neuropsychological batteries and other samples.

Keywords: ADNI, factorial invariance, neuropsychological battery, mild cognitive impairment, Alzheimer's disease

Supplemental materials: https://doi.org/10.1037/neu0000736.supp

M. Chang https://orcid.org/0000-0002-7026-4197

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). As such, the investigators with the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Correspondence concerning this article should be addressed to M. Chang, Department of Psychology and Human Neuroscience Institute, G341 MVR Hall, Cornell University, Ithaca, NY 14853, United States. Email: mc2674@cornell.edu

Alzheimer's disease (AD) is a neurodegenerative brain condition that causes gradual loss of cognitive functions (especially memory and language), and eventually, death. Mild cognitive impairment (MCI) is AD's prodromal stage, during which individuals have clinically significant impairment in one of the five cognitive domains that figure in dementia diagnoses, usually episodic memory (Petersen, 2004, 2011). As life expectancy has increased in the United States, the death rate from AD has increased 146% between 2000 and 2018, making it the fifth leading cause of mortality in adults aged 65 or older (Alzheimer's Association, 2020). One of the priorities of AD research is to develop methods of diagnosing it at its earliest stages, when interventions can be most effective (Mueller et al., 2005b). Here, considerable effort has been devoted to identifying cognitive, biological and neuroimaging markers of AD, with the Alzheimer's Disease Neuroimaging Initiative (ADNI) being one of the most extensive projects of this sort (Mueller et al., 2005a; Weiner & Veitch, 2015).

The ADNI is an ongoing, longitudinal, multicenter study that aims at advancing knowledge about diagnosis and progression of AD. The original 5-year study, ADNI 1, recruited large samples of healthy controls (HC), MCI, and AD subjects and evaluated changes in cognitive, biological, and neuroimaging markers over multiple sessions. Cognitive markers were measured with a comprehensive neuropsychological battery, which incorporated gold-standard assessments for multiple domains, including episodic memory, language, attention, visuospatial processing, and executive function. Since the ADNI's inception, several analyses of data from portions of this neuropsychological battery have been published (e.g., Brainerd et al., 2014; Crane et al., 2012; Gibbons et al., 2012; Johnson et al., 2012; Park et al., 2012; Petersen et al., 2010).

A common practice in analyzing the ADNI battery data is to compare the instrument scores, composite scores (i.e., the average of multiple instrument scores), or factor scores (i.e., the estimated values of latent factors) between diagnostic groups (e.g., Crane et al., 2012; Giraldo et al., 2017; Petersen et al., 2010). In such work, researchers implicitly assume that the neuropsychological tests are measuring the same underlying cognitive functions in the same way, across different diagnostic groups. Psychometrically, such assumptions are referred to as measurement invariance. If measurement invariance does not hold, a neuropsychological battery can be biased against particular diagnostic groups, which renders between-group comparisons invalid (Wu et al., 2007). However, despite the fundamental importance of measurement invariance, it has received only limited attention to date (Meredith & Teresi, 2006; Vandenberg & Lance, 2000).

Normally, measurement invariance is evaluated with multigroup confirmatory factor analysis (MG-CFA; Mungas et al., 2011; Wu et al., 2007). This approach views measurement invariance as a form of factorial invariance. To test factorial invariance, researchers place equality restrictions on a series of multigroup models and test model fits. Such analyses determine whether a neuropsychological battery is assessing the same latent factors across groups and whether the associations between test items and latent factors remain invariant across groups. The results of these invariance tests, in turn, decide whether valid comparisons can be made between groups. We will discuss some further details of factorial invariance tests in the Method section.

Factorial Invariance Across Diagnostic Groups in Cognitive Aging

In the cognitive aging literature, considerable research has been devoted to establishing factorial invariance of neuropsychological batteries among groups of different ages or genders (de Frias & Dixon, 2005; Dowling et al., 2010; Rawlings et al., 2016) and groups with different natural languages or ethnicities (Flores et al., 2017; Mungas et al., 2011; Pedraza et al., 2005; Siedlecki et al., 2010; Tuokko et al., 2009). However, data on factorial invariance among groups of older adults with different degrees of cognitive impairment are quite thin and mixed, considering their clinical importance. As discussed below, some studies suggest that factor structures vary with clinical diagnoses (Delis et al., 2003; Jones & Ayers, 2006; Kanne et al., 1998; Siedlecki et al., 2008), whereas others favor factorial invariance (Hayden et al., 2011; Johnson et al., 2008; Mitchell et al., 2012; Park et al., 2012).

Evidence for Factorial Instability

Kanne et al. (1998) applied principal components analyses (PCA) to data generated by the neuropsychological battery of the Alzheimer's Disease Research Center (ADRC). They found that although a one-factor model accounted for the data of an HC group, the data of a very mild AD group and a mild AD group required a three-factor model (verbal memory, visuospatial processing, executive control). Similarly, Jones and Ayers (2006) conducted exploratory factor analyses (EFA) of data from the expanded Consortium to Establish a Registry for Alzheimer's Disease (CERAD) battery. Their results yielded a one-factor solution for a demented group but a two-factor solution (dementia severity, memory) for a mixed group of demented and nondemented subjects. Moreover, Delis et al. (2003) found that the California Verbal Learning Test (CVLT), an episodic memory battery, yielded a one-factor solution for an HC group, but a two-factor solution for an AD group. The difference lay in whether immediate and delayed memory tests loaded on the same or separate factors. Thus, these three investigations showed that the neuropsychological performance of demented and nondemented groups did not share a common factor structure, indicating that some instruments may tap different cognitive functions in different groups.

Although these studies all argue against factorial invariance, none of them formally evaluated factorial invariance with MG-CFA. Consequently, although the studies revealed different factor structures for different groups, they did not establish that between-group factor differences were statistically reliable. Siedlecki et al. (2008) filled this gap by conducting MG-CFA for the ADRC battery, comparing HC, questionable dementia (QD), and probable AD groups. In addition, they also conducted EFA for each diagnostic group. The MG-CFA results revealed that the factor structures were significantly different across the three diagnostic groups. The EFA results showed that the key difference was that the probable AD group exhibited two episodic memory factors (immediate and delayed memory), whereas the other two groups exhibited only one.

Evidence for Factorial Invariance

In contrast to the above studies, Johnson et al. (2008) reported that a model with a general cognitive factor and three specific factors (verbal memory, working memory, visuospatial processing)

accounted for the data of a 12-item neuropsychological battery in both HC and demented groups. However, this battery did not contain tests for other cognitive domains that are routinely included in neuropsychological batteries, such as executive function and attention.

Using a more comprehensive battery (from the National Alzheimer's Disease Coordinating Center [NACC]), Hayden et al. (2011) identified a four-factor model (memory, attention, executive function, language) that was invariant across HC, MCI, and demented groups. In addition, that structure remained stable over a 1-year interval. Soon thereafter, Mitchell et al. (2012) reported a similar pattern with the same battery. They identified a four-factor model (memory/language, processing speed/executive function, attention, cognitive reserve), and it proved to be invariant across two groups (HC and mixed amnestic-MCI/AD).

In addition, Park et al. (2012) reported an MG-CFA of the baseline data for the ADNI neuropsychological battery, the battery that figures in the current article. Their analysis was focused on an a priori model in which five distinct factors reflect the specific cognitive functions that the battery is assumed to tap (memory, language, visuospatial processing, attention, executive function). Their MG-CFA results supported this five-factor model and showed that it was invariant between the less and more functionally impaired groups.

The Present Study

As we have just seen, available findings on whether the factor structure of neuropsychological batteries is invariant across diagnostic groups in older adults are neither extensive nor consistent. Moreover, as far as we are aware, there is only one study examining factorial invariance for the ADNI battery (Park et al., 2012), which is obviously insufficient considering the ADNI's substantial impact on cognitive aging research (Weiner et al., 2015, 2017).

In the current article, we report an expansion of work on factorial invariance for the ADNI battery. The ADNI subject pool consists of three diagnostic groups (HC, MCI, and AD), each of which contains a large number of subjects. After baseline testing in the ADNI 1, follow-up testing occurred at 6-month intervals for 5 years, with the full neuropsychological battery being readministered on each occasion. We analyzed data from the first 2 years of testing, during which subject samples remained large, in order to determine whether the underlying factor structure of the battery remained stable across diagnostic groups and over time. We first conducted EFAs for each diagnostic group in each testing session, which allowed us to formalize the core cognitive abilities that the battery taps. Then, we evaluated the results of the EFAs with CFAs by fitting models to each group within each session, which is a prerequisite for evaluating multigroup invariance (Brown, 2015). Finally, we tested factorial invariance with a series of MG-CFAs, both between the three diagnostic groups and across the 2 years of testing.

Method

Subjects

Data analyzed in this article were obtained from the ADNI database (http://adni.loni.usc.edu/). Because full details of the ADNI 1 subject sample have been provided in several prior publications (e.g., Weiner & Veitch, 2015), we only report key demographic characteristics in Table 1, without elaboration. We originally planned to include five distinct testing sessions (baseline,

Table 1
Demographic Characteristics for ADNI Subjects

Session	Characteristics	Group		
		HC	MCI	AD
Baseline				
	Sample size	206	364	178
	Age	75.89 (4.96)	74.84 (7.25)	75.6 (7.45)
	Education	16.1 (2.92)	15.7 (2.96)	14.85 (3.09)
	Gender, female (%)	49.03%	34.89%	48.88%
	APOE ε4 carriers (%)	23.79%	55.49%	65.17%
6 months				
	Sample size	204	325	188
	Age	76.39 (5.23)	75.46 (7.16)	75.71 (7.45)
	Education	16.08 (2.91)	15.69 (2.95)	14.94 (3.08)
	Gender, female (%)	47.55%	35.08%	47.34%
	APOE ε4 carriers (%)	24.02%	56.62%	65.96%
12 months				
	Sample size	197	269	213
	Age	76.75 (5.34)	76.3 (7.14)	76.14 (7.29)
	Education	16.15 (2.81)	15.88 (2.94)	15.02 (3.04)
	Gender, female (%)	48.73%	34.57%	45.07%
	APOE ε4 carriers (%)	23.35%	55.39%	65.26%
24 months				
	Sample size	191	169	234
	Age	77.9 (5.43)	76.9 (6.88)	77.1 (7.3)
	Education	16.18 (2.81)	15.89 (2.87)	15.03 (3.11)
	Gender, female (%)	46.60%	34.91%	44.44%
	APOE ε4 carriers (%)	25.65%	48.52%	69.23%

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative; HC = healthy control; MCI = mild cognitive impairment; AD = Alzheimer's disease. Standard deviations are shown in parentheses.

6 months, 12 months, 18 months, 24 months) in our analyses. The neuropsychological battery was administered in full during each session. However, a preliminary review of the data revealed that substantially fewer subjects were tested during the 18-month session than during the other four sessions. Consequently, the analyses were confined to data from the baseline, 6-, 12-, and 24-month sessions.

Clinical diagnoses of individual subjects were established at baseline and were also reestablished during each subsequent session (Mueller et al., 2005b). Subjects in the HC, MCI, and AD groups were carefully screened and classified, according to five criteria: (a) memory impairment: MCI and AD subjects must have memory complaints whereas HC subjects must not; (b) education-adjusted scores for the Logical Memory II subscale of the Wechsler Memory Scale-Revised (Wechsler, 1987): for ≥16 years of education, HC subjects must have scores ≥9, MCI subjects 9–11, AD subjects ≤8; for 8–15 years of education, HC subjects must have scores ≥5, MCI subjects 5–9, AD subjects ≤4; for 0–7 years of education, HC subjects must have scores ≥3, MCI subjects 3–6, AD subjects ≤2; (c) Mini Mental State Exam scores: HC and MCI subjects must have scores between 24 and 30, and AD subjects must have scores between 20 and 26; (d) Clinical Dementia Rating (CDR; Morris, 1993) scores: HC subjects must have both total score and memory box score = 0; MCI subjects must have both total score and memory box score = 0.5 (This means that the MCI sample is overwhelmingly amnestic-MCI, which is the most common subtype); AD subjects must have total score = 0.5 or 1; (e) general cognition and functioning: HC subjects must have no significant impairment in normal cognitive functions or daily living; MCI subjects must have relatively preserved general cognitive and functional performance that the site physicians cannot make an AD diagnosis; and AD subjects must have met the National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD. Full details of the screening criteria can be found in the ADNI 1 Procedures Manual (http://adni.loni.usc.edu/methods/documents/).

Measures

Complete details of the neuropsychological battery and assessment methods can be found in the ADNI 1 Procedures Manual. Sixteen variables of interest (from seven neuropsychological tests) were selected for the factor analyses, as described below. The descriptive statistics for the 16 variables across the three diagnostic groups (HC, MCI, AD) over the four sessions (baseline, 6 months, 12 months, 24 months) are summarized in Table 2.

Episodic Memory Tests

Alzheimer's Disease Assessment Scale—Cognitive (ADAS-Cog), Delayed Recall and Recognition. ADAS-Cog (Rosen et al., 1984) has been widely used in clinical trials for AD. In the word recall task, subjects are presented with 10 high-frequency high-imagery nouns and are asked to recall as many as they can. This is followed by two additional trials of study and recall for the same word list. After the three study-recall trials, subjects complete two other subtests of the ADAS-Cog, followed by a delayed recall test for the 10 words. Subsequently, subjects complete three other subtests of ADAS-Cog and are then administered a recognition test. In this task, subjects first study 12 words, followed by an old–new recognition test composed of 24 words (the 12 old ones + 12 new ones). We included the percentage of correct delayed recall and the percentage of correct recognition in our factor analyses. We did not include immediate recall, because it was too strongly correlated with delayed recall for the two measures to pass discriminant validity tests, which can cause incorrect factor analyses. We originally considered using the recall decline between the third immediate test and the delayed test as an indicator of forgetting, but that measure produced too many negative values to be practicable.

Rey Auditory Verbal Learning Test, Recall and Forgetting. The Rey Auditory Verbal Learning Test (RAVLT; Rey, 1964) is the classic clinical episodic memory test. Subjects first complete five study-recall trials for a single 15-word list (List A). Next, they study a new 15-word list (List B) and complete an immediate free recall for List B. After that, the subjects are asked to again recall List A. Finally, after a 30 min delay, subjects are asked to recall List A again. For the current study, we used the average percentage of correct recall on the first five trials as an indicator of recall accuracy, and the percentage of recall decline between the fifth recall test and the delayed test after 30 min as an indicator of forgetting.

Visuospatial Processing Tests

Clock Drawing Test, Command and Copy. Clock Drawing Test (CDT) is designed to measure construction abilities, and it consists of two parts: command and copy. In the command part, subjects follow the test administrator's instruction to draw a clock, such as "draw the face of a clock showing the numbers and two hands set to ten after eleven." In the copy part, subjects are presented with a clock printed on a response sheet and are instructed to copy it. CDT scores take into account approximate circular shape, symmetry of number placement, presence of hands, and accuracy of number and hand placement. We included both the CDT command and copy scores in our analysis.

ADAS-Cog, Construction Praxis Test. This test measures subjects' ability to copy four geometric figures: a circle, a pair of overlapping rectangles, a diamond (rhombus), and a cube. The figures are presented individually, in the order just described. A subject's score is simply the total number of figures correctly copied, and we used that score in our analyses.

Executive Function Tests

Trail Making Test, Parts A and B. The Trail Making Test (TMT; Reitan & Wolfson, 1985) is a test of processing speed and executive function. It consists of two parts. In Part A, subjects are required to draw a line to connect a series of numbers in ascending numerical order. In Part B, they are asked to draw a line to connect a series of numbers and letters alternately in ascending numerical and alphabetical order. Part A is meant to test visuomotor and visual scanning abilities, and Part B is meant to test these two abilities plus cognitive shifting. TMT scores are the total time used to complete each part, with a maximum of 150 s allowed for Part A and a maximum of 300 s allowed for Part B. We included both the Part A and the Part B scores in our analyses.

ADAS-Cog, Number Cancellation Test. This is a test for visual attention and processing speed, in which subjects are instructed to cross out two designated letters within several lines of

Table 2
Descriptive Statistics of the ADNI Neuropsychological Battery

Sessions	Assessments	HC	MCI	AD	*p* value	$p < .05^*$
Baseline						
	ADAS delayed recall	.72 (.17)	.38 (.23)	.13 (.15)	<.001	a, b, c
	ADAS recognition	.79 (.19)	.62 (.22)	.45 (.23)	<.001	a, b, c
	RAVLT recall	.58 (.12)	.41 (.12)	.31 (.10)	<.001	a, b, c
	RAVLT forgetting	.35 (.26)	.67 (.31)	.88 (.23)	<.001	a, b, c
	CDT command	4.68 (.66)	4.18 (.99)	3.35 (1.32)	<.001	a, b, c
	CDT copy	4.88 (.39)	4.63 (.70)	4.34 (.99)	<.001	a, b, c
	ADAS construction	4.63 (.49)	4.45 (.57)	4.13 (.68)	<.001	a, b, c
	Digit span forward	8.75 (1.97)	8.21 (1.96)	7.55 (1.88)	<.001	a, b, c
	Digit span backward	7.18 (2.16)	6.14 (2.03)	5.03 (1.83)	<.001	a, b, c
	CFT animal	20.00 (5.57)	16.00 (4.86)	12.39 (5.00)	<.001	a, b, c
	CFT vegetable	14.89 (3.88)	10.76 (3.39)	7.81 (3.36)	<.001	a, b, c
	Boston naming test	27.83 (2.34)	25.56 (4.04)	22.13 (6.30)	<.001	a, b, c
	ADAS naming	4.93 (.26)	4.74 (.50)	4.46 (.80)	<.001	a, b, c
	TMT Part A	36.88 (13.11)	44.01 (21.26)	67.19 (34.87)	<.001	a, b, c
	TMT Part B	89.37 (42.40)	128.04 (68.39)	193.52 (81.07)	<.001	a, b, c
	ADAS cancellation	4.57 (.62)	4.04 (.98)	3.15 (1.33)	<.001	a, b, c
6 months						
	ADAS delayed recall	.73 (.17)	.37 (.24)	.13 (.17)	<.001	a, b, c
	ADAS recognition	.81 (.17)	.61 (.25)	.38 (.24)	<.001	a, b, c
	RAVLT recall	.56 (.13)	.38 (.12)	.27 (.10)	<.001	a, b, c
	RAVLT forgetting	.36 (.25)	.73 (.29)	.93 (.18)	<.001	a, b, c
	CDT command	4.64 (.62)	4.14 (1.05)	3.28 (1.39)	<.001	a, b, c
	CDT copy	4.85 (.36)	4.64 (.66)	4.18 (1.16)	<.001	a, b, c
	ADAS construction	4.61 (.53)	4.45 (.59)	4.10 (.78)	<.001	a, b, c
	Digit span forward	8.80 (1.91)	8.06 (2.04)	7.36 (2.02)	<.001	a, b, c
	Digit span backward	7.18 (2.24)	6.00 (1.93)	5.08 (1.98)	<.001	a, b, c
	CFT animal	20.35 (5.61)	15.66 (4.85)	11.72 (4.60)	<.001	a, b, c
	CFT vegetable	14.32 (4.21)	10.44 (3.74)	7.55 (3.56)	<.001	a, b, c
	Boston naming test	28.22 (2.08)	25.55 (4.60)	22.11 (6.26)	<.001	a, b, c
	ADAS naming	4.95 (.25)	4.73 (.51)	4.43 (.81)	<.001	a, b, c
	TMT Part A	33.86 (12.04)	44.45 (19.93)	67.59 (37.51)	<.001	a, b, c
	TMT Part B	84.26 (38.77)	129.28 (70.07)	202.47 (80.81)	<.001	a, b, c
	ADAS cancellation	4.38 (.72)	3.94 (.98)	3.02 (1.34)	<.001	a, b, c
12 months						
	ADAS delayed recall	.72 (.19)	.37 (.24)	.12 (.15)	<.001	a, b, c
	ADAS recognition	.84 (.15)	.64 (.26)	.39 (.25)	<.001	a, b, c
	RAVLT recall	.59 (.14)	.41 (.13)	.28 (.11)	<.001	a, b, c
	RAVLT forgetting	.33 (.26)	.63 (.33)	.86 (.24)	<.001	a, b, c
	CDT command	4.69 (.61)	4.17 (1.03)	3.25 (1.39)	<.001	a, b, c
	CDT copy	4.85 (.41)	4.65 (.66)	4.12 (1.21)	<.001	a, b, c
	ADAS construction	4.6 (.57)	4.46 (.62)	4.13 (.81)	<.001	a, b, c
	Digit span forward	8.84 (1.93)	8.09 (2.10)	7.25 (2.07)	<.001	a, b, c
	Digit span backward	7.31 (2.32)	6.05 (2.07)	4.94 (1.92)	<.001	a, b, c
	CFT animal	20.69 (5.35)	15.88 (5.73)	11.55 (4.85)	<.001	a, b, c
	CFT vegetable	14.76 (4.26)	10.73 (3.85)	7.17 (3.53)	<.001	a, b, c
	Boston naming test	28.43 (1.92)	25.82 (4.92)	21.57 (6.48)	<.001	a, b, c
	ADAS naming	4.96 (.32)	4.72 (.59)	4.41 (.81)	<.001	a, b, c
	TMT Part A	34.23 (10.11)	43.64 (21.03)	64.39 (35.08)	<.001	a, b, c
	TMT Part B	80.4 (36.6)	127.8 (71.51)	197.84 (86.17)	<.001	a, b, c
	ADAS cancellation	4.59 (.83)	4.18 (1.03)	3.08 (1.54)	<.001	a, b, c
24 months						
	ADAS delayed recall	.73 (.18)	.38 (.24)	.12 (.17)	<.001	a, b, c
	ADAS recognition	.81 (.18)	.61 (.23)	.36 (.25)	<.001	a, b, c
	RAVLT recall	.60 (.14)	.42 (.13)	.26 (.12)	<.001	a, b, c
	RAVLT forgetting	.31 (.28)	.62 (.31)	.82 (.25)	<.001	a, b, c
	CDT command	4.71 (.62)	4.36 (.80)	3.02 (1.42)	<.001	a, b, c
	CDT copy	4.85 (.38)	4.65 (.68)	3.95 (1.34)	<.001	a, b, c
	ADAS construction	4.68 (.47)	4.54 (.58)	3.94 (1.00)	<.001	a, b, c
	Digit span forward	8.94 (2.02)	8.07 (2.04)	6.85 (2.29)	<.001	a, b, c
	Digit span backward	7.51 (2.24)	6.41 (2.14)	4.70 (2.04)	<.001	a, b, c
	CFT animal	20.88 (5.67)	15.78 (4.95)	10.59 (5.52)	<.001	a, b, c
	CFT vegetable	14.83 (4.20)	10.78 (3.95)	6.42 (3.97)	<.001	a, b, c

Table 2 (*continued*)

Sessions	Assessments	HC	MCI	AD	*p* value	$p < .05^*$
	Boston naming test	28.42 (2.21)	26.11 (4.47)	20.64 (7.49)	<.001	a, b, c
	ADAS naming	4.95 (.21)	4.75 (.59)	4.15 (1.14)	<.001	a, b, c
	TMT Part A	32.66 (10.82)	42.82 (24.41)	67.95 (38.30)	<.001	a, b, c
	TMT Part B	83.92 (42.45)	116.60 (63.83)	204.03 (89.28)	<.001	a, b, c
	ADAS cancellation	4.37 (.82)	3.93 (1.02)	2.65 (1.51)	<.001	a, b, c

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative; HC = healthy control; MCI = mild cognitive impairment; AD = Alzheimer's disease; ADAS = Alzheimer's Disease Assessment Scale; RAVLT = Rey Auditory Verbal Learning test; CDT = Clock Drawing Test; CFT = Categorical Fluency Test; TMT = Trail Making Test; The *p* value column indicates whether there is significant difference among the three groups.
* Multiple comparison abbreviations: a = HC differs from MCI; b = MCI differs from AD; c = HC differs from AD.

mixed letters. The maximum time allowed was 45 s. We included the cancellation test score in our analyses, which is the number of letters correctly crossed out minus the number of letters incorrectly crossed out. The test score was transformed into a 5-point scale.

Attention Tests

Digit Span, Forward, and Backward. Digit span is a test for attention and working memory. In the forward part, the administrator reads five numbers to each subject, and the subject is required to repeat the numbers in the same order. In the backward part, the administrator reads three numbers to each subject, and the subject is required to repeat the numbers in reverse order. We used both total correct forward and backward span in our factor analyses.

Language Tests

Category Fluency Test (CFT), Animals and Vegetables. The CFT (Harrison et al., 2000) is a measure of semantic memory. In two separate 20-s trials, subjects are required to generate exemplars of a given semantic category (animal or vegetable). In each instance, the test score is the number of correct unique exemplars named. The total score for animals and the total score for vegetables were used in our factor analyses.

Boston Naming Test (BNT). This is a test of object recognition and naming (Goodglass et al., 1983). Subjects are presented with 30 black-and-white line drawings, and they are required to name the object in each drawing. The drawings are ordered from the most common (bed) to the least common (protractor). Subjects are given a maximum of 20 s for each response. If subjects' responses indicate misperception of a drawing, a semantic cue is given. If subjects' responses are still incorrect, a phonetic cue (the first phoneme of the object's name) is given. The BNT score that we used was the total number of correct responses, regardless of whether or not they were cued.

ADAS-Cog, Naming Test. In this test, subjects are instructed to name 12 randomly selected objects with high, medium and low frequency values. They are also required to name the fingers in their dominant hand. The ADAS naming score is simply the total number of correct responses, converted into a 5-point scale.

Statistical Analysis

We used both the exploratory (EFA) and confirmatory (CFA) approaches to identify the latent structure underlying the ADNI battery and to evaluate its stability across diagnostic groups and over time. All of our analyses were conducted in R version 3.6.1. We performed the EFAs with the psych package (Revelle, 2016) and the CFAs with the lavaan package (Hirschfeld & von Brachel, 2014; Rosseel, 2012). The five tests from the ADAS-Cog (delayed recall, recognition, construction, number cancellation, and naming) were all reverse-coded compared to the ADAS-Cog grading manual. This ensures that higher scores indicate better performance, which reduces convergence problems (Gustafsson & Stahl, 2005). We also winsorized the top and bottom 5% of the two timed tests (TMT Parts A and B) to eliminate outliers, which helped minimize Heywood case problems (Yuan & Bentler, 2001).

EFA Analyses

An EFA was conducted for each diagnostic group (HC, MCI, AD) in each of the four sessions (baseline, 6 months, 12 months, 24 months), yielding 12 EFAs. These analyses determined the model that best captured neuropsychological performance in each data subset. For all EFAs, we followed the same procedure. To begin with, we used the Kaiser-Meyer-Olkin (KMO; Kaiser, 1970; Kaiser & Rice, 1974) measure of sampling adequacy to determine which variables should be retained in the factor analyses. We implemented the standard criterion that only variables with a KMO score >.50 are suitable for factor analysis (Williams et al., 2010; Yong & Pearce, 2013). Next, we computed Bartlett's test of sphericity (Bartlett, 1950) to confirm that the correlation matrix is significantly different from an identity matrix, which is a prerequisite for conducting a factor analysis. The null hypothesis that the correlation matrix is not significantly different from an identity matrix was rejected at the .05 level of confidence. Additionally, we verified that the determinant of the correlation matrix was >.00001, which ruled out multicollinearity problems.

When determining how many factors to extract, we closely followed Costello and Osborne's (2005) recommendations. First, we ran multiple factor analyses with factor numbers chosen based on Kaiser's criterion (Kaiser, 1958), scree plots (Cattell & Vogelmann, 1977), parallel analysis (Horn, 1965) and *a priori* factor structure. As these approaches often suggested different numbers of factors, we also ran factor analyses with the number of factors set at numbers above and below these suggested numbers. Finally, we chose the best-fitting factor solution based on the following criteria: (a) statistical interpretation (i.e., simple structure); (b) theoretical significance (i.e., consistency with the prior literature); (c) parsimony (i.e., smallest number of factors); and (d) absence of Heywood cases, because Heywood cases (commonality >1, negative residual variances; Heywood, 1931) are usually indicators of model

misspecification (Brown, 2015; Loehlin & Beaujean, 2016). We chose oblimin rotations as there were weak to moderate correlations among factors (Costello & Osborne, 2005). As for factor extraction methods, we used principal axis factoring because the multivariate normality assumption was not met (Costello & Osborne, 2005).

CFA Analyses

Following Brown's (2015) recommendations, we first conducted single group CFA for each diagnostic group in each of the four sessions, yielding 12 CFAs. In these CFAs, we tested fits for both the data-driven and the theory-driven models. The former are the factor solutions produced by the EFAs, and the latter is a five-factor model (memory, visuospatial processing, attention, language, executive function) proposed in prior research with the ADNI baseline data (Johnson et al., 2012; Park et al., 2012). We originally planned to pit these models against each other and pick the best-fitting one for subsequent factorial invariance analyses. However, as will be seen below, that was unnecessary owing to the high agreement between the data-driven and theory-driven models.

After establishing fit for each group within each session, we proceeded to the MG-CFAs. We first determined whether the factor structure remained stable across each pair of diagnostic groups (HC vs. MCI, MCI vs. AD, and HC vs. AD) within each testing session. Second, we evaluated whether each diagnostic group's factor structure was invariant over the 24-month interval. Here, we only considered the clinical diagnoses with respect to each session. In other words, instead of examining the trajectory of each individual's clinical status over time, we focused on the factor structure of the overall diagnostic group in each session.

In MG-CFAs, we examined four increasingly rigorous levels of factorial invariance: configural, weak, strong, and strict invariance. The four levels of invariance are formulated with a nested hierarchy of models that were imposed with increasingly stringent equality restrictions (Beaujean, 2014; Brown, 2015; Cudeck & MacCallum, 2007; Wu et al., 2007). The most basic level of invariance, *configural invariance* or *equal form*, requires that the number of factors is the same across groups, and the same items load on the same factors in all groups. This level of invariance provides strong support for the hypothesis that the neuropsychological battery measures the same cognitive functions in all diagnostic groups. Configural invariance is a prerequisite to test subsequent levels of invariance. *Weak invariance* or *equal factor loadings* requires that, in addition to the identical model configuration, the factor loadings are equal across groups, which means that the regression slopes between the underlying factor scores and the individual test scores are identical in all groups. This level of invariance must be present in order to compare the factor variances or covariances across groups. If weak invariance is rejected, it is meaningless to proceed to test the other two levels of factorial invariance. *Strong invariance* or *equal intercepts* requires that the factor loadings *and* their intercepts are invariant across groups, which means that in addition to equal regression slopes, mean test scores are the same across groups when factor scores are zero. This level of invariance is necessary for making inferences about group differences in latent factor means based on group differences in test scores. The failure to establish strong invariance rules out strict invariance. Last, *strict invariance* requires that factor loadings, intercepts, and residual variances are all equal; that is, in addition to strong invariance, any variance in test scores that is not explained by common factors must be equal across groups. This indicates that the reliabilities of the individual tests are equal across groups, and thus any group differences in test scores are entirely due to differences in the underlying cognitive functions. In the present study, we placed greater emphasis on the interpretation of configural, weak, and strong invariance, although we still report statistics for strict invariance when the prior three levels of invariance are all achieved.

Model fits are traditionally examined with maximum likelihood tests. Absolute levels of fit are indicated by χ^2 values, and comparative fits between two nested models are indicated by differences in χ^2 values ($\Delta\chi^2$; Satorra & Bentler, 2001). However, both $\chi^2$and $\Delta\chi^2$ have been criticized for being sensitive to differences in sample size (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). To overcome this limitation, Bentler's comparative fit index (CFI) and root mean square of approximation (RMSEA) were developed as alternative fit indexes and have been commonly used in recent years. CFI is a goodness-of-fit index with higher values indicating better fits. It compares fit of a hypothesized model to the fit of a simpler model while adjusting for model complexity or parsimony (Iacobucci, 2010). RMSEA is a "badness-of-fit" index with lower values indicating better fits, and it is relatively robust against variation in sample size and distribution (Steiger, 1990). Conventionally, CFI $\geq$ .95 and RMSEA $\leq$ .06 are used as cutoffs for excellent fit, and CFI $\geq$ .90 and RMSEA $\leq$ .08 are used as cutoffs for adequate fit (Browne & Cudek, 1993; Hu & Bentler, 1999; Jöreskog & Sörbom, 1993). We relied primarily on CFI and RMSEA for model fit decisions for two reasons. First, as discussed above, they have obvious advantages compared to more traditional indexes, and they have demonstrated good performance overall (Hu & Bentler, 1998). Second, they are the most commonly used fit indexes in prior measurement invariance studies for aging data sets (e.g., Hayden et al., 2011; Johnson et al., 2008; Park et al., 2012; Siedlecki et al., 2008), which facilities comparison of our results to earlier findings. In addition, we report the ΔCFI between adjacent levels of invariance, where ΔCFI > .01 has been used to indicate significant differences in fits (Cheung & Rensvold, 2002).

Results

Descriptive statistics for all the neuropsychological instruments are reported separately for the HC, MCI, and AD groups and for the four sessions in Table 2. Among the 16 instruments that were included in the factor analyses, the HC group performed significantly better than the MCI and AD groups, and the MCI group performed significantly better than the AD group.

EFA Results

Results of the EFA analyses are summarized in Table 3. Following the usual convention in factor analysis, we treated loadings $\geq$.40 as the cutoff for significant loadings (Lindeman, 1980; Nunnally, 1994), and thus, we only display variables with loadings $\geq$.40 in Table 3. Visual inspection of Table 3 reveals two notable patterns. First, the factor solutions are highly interpretable, which indicates that the neuropsychological instruments are measuring the expected cognitive functions. Despite slight deviations among the three diagnostic groups and across the four sessions, we see that (a) the episodic memory instruments (ADAS delayed recall, ADAS

Table 3

Exploratory Factor Analysis Results of the ADNI Neuropsychological Battery for the Three Diagnostic Groups Across the Four Sessions

HC		MCI		AD	
Baseline					
Factor 1 (Memory)		Factor 1 (Memory)		Factor 1 (Executive/visuospatial)	
ADAS delayed recall	.68	ADAS delayed recall	.74	CDT command	.66
ADAS recognition	.42	ADAS recognition	.57	CDT copy	.70
RAVLT recall	.75	RAVLT recall	.65	ADAS construction	.71
RAVLT forgetting	−.60	RAVLT forgetting	−.71	TMT Part A	−.59
				ADAS cancellation	.44
Factor 2 (Executive function)		Factor 2 (Language)		Factor 2 (Language)	
TMT Part A	.69	CFT animal	.57	CFT animal	.59
TMT Part B	.77	CFT vegetable	.47	CFT vegetable	.49
		Boston naming test	.82	Boston naming test	.74
		ADAS naming	.53	ADAS naming	.81
Factor 3 (Attention)		Factor 3 (Executive function)		Factor 3 (Memory)	
Digit span forward	.83	TMT Part A	−.72	ADAS delayed recall	.68
Digit span backward	.73	TMT Part B	.62	ADAS recognition	.54
		ADAS cancellation	.60	RAVLT recall	.63
				RAVLT forgetting	−.44
Factor 4 (Visuospatial processing)		Factor 4 (Attention)		Factor 4 (Attention)	
CDT command	.62	Digit span forward	.70	Digit span forward	.55
CDT copy	.76	Digit span backward	.67	Digit span backward	.52
Factor 5 (Language)		Factor 5 (Visuospatial processing)			
CFT animal	.46	CDT command	.69		
CFT vegetable	.49	CDT copy	.53		
6 months					
Factor 1 (Memory)		Factor 1 (Executive/visuospatial)		Factor 1 (Executive/visuospatial)	
ADAS delayed recall	.73	CDT command	.60	CDT command	.76
ADAS recognition	.41	CDT copy	.56	CDT copy	.75
RAVLT recall	.80	ADAS construction	.45	ADAS construction	.53
RAVLT forgetting	−.56	TMT Part A	−.58	TMT Part A	−.67
		TMT Part B	−.62	TMT Part B	−.43
		ADAS cancellation	.51	ADAS cancellation	.73
				Digit span forward	.48
Factor 2 (Executive function)		Factor 2 (Memory)		Factor 2 (Language)	
TMT Part A	.72	ADAS delayed recall	.86	CFT animal	.45
TMT Part B	.81	ADAS recognition	.50	CFT vegetable	.45
		RAVLT recall	.59	Boston naming test	.87
		RAVLT forgetting	−.70	ADAS naming	.72
Factor 3 (Attention)		Factor 3 (Language)		Factor 3 (Memory)	
Digit span forward	.67	CFT animal	.66	ADAS delayed recall	.60
Digit span backward	.90	CFT vegetable	.59	ADAS recognition	.64
		Boston naming test	.68	RAVLT recall	.48
		ADAS naming	.50	RAVLT forgetting	−.65
Factor 4 (Language)		Factor 4 (Attention)		Factor 4 (Attention)	
CFT animal	.67	Digit span forward	1.00	Digit span forward	.81
CFT vegetable	.46	Digit span backward	.47	Digit span backward	.41
Boston naming test	.46				
Factor 5 (Visuospatial processing)					
CDT command	.46				
CDT copy	.59				
12 months					
Factor 1 (Memory)		Factor 1 (Executive/visuospatial)		Factor 1 (Executive/visuospatial)	
ADAS delayed recall	.82	CDT command	.44	CDT command	.67
ADAS recognition	.45	CDT copy	.55	CDT copy	.66
RAVLT recall	.76	ADAS construction	.45	ADAS construction	.63
RAVLT forgetting	−.41	TMT Part A	−.75	TMT Part A	−.80
		TMT Part B	−.74	TMT Part B	−.45
		ADAS cancellation	.57	ADAS cancellation	.77
Factor 2 (Visuospatial processing)		Factor 2 (Memory)		Factor 2 (Language)	
CDT command	.81	ADAS delayed recall	.88	Boston naming test	.76
CDT copy	.64	ADAS recognition	.55	ADAS naming	.85
		RAVLT recall	.68		
		RAVLT forgetting	−.66		

(*table continues*)

Table 3 (*continued*)

HC		MCI		AD	
Factor 3 (Attention)		Factor 3 (Language)		Factor 3 (Memory)	
Digit span forward	.81	CFT animal	.52	ADAS delayed recall	.57
Digit span backward	.70	CFT vegetable	.56	ADAS recognition	.42
		Boston naming test	.76	RAVLT recall	.44
		ADAS naming	.75	RAVLT forgetting	−.40
Factor 4 (Executive function)		Factor 4 (Attention)		Factor 4 (Attention)	
TMT Part A	.76	Digit span forward	.86	Digit span forward	.62
TMT Part B	.66	Digit span backward	.51	Digit span backward	.69
Factor 5 (Language)					
Boston naming test	.48				
CFT animal	.72				
24 months					
Factor 1 (Memory)		Factor 1 (Memory)		Factor 1 (Executive/visuospatial)	
ADAS delayed recall	.78	ADAS delayed recall	.85	CDT command	.51
ADAS recognition	.56	ADAS recognition	.53	CDT copy	.74
RAVLT recall	.65	RAVLT recall	.77	ADAS construction	.71
		RAVLT forgetting	−.68	TMT Part A	−.88
				ADAS cancellation	.74
				Digit span backward	.42
Factor 2 (Attention)		Factor 2 (Visuospatial processing)		Factor 2 (Language)	
Digit span forward	.98	CDT command	.48	CFT animal	.53
Digit span backward	.53	CDT copy	.47	CFT vegetable	.58
				Boston naming test	.91
				ADAS naming	.79
Factor 3 (Executive function)		Factor 3 (Language)		Factor 3 (Attention)	
TMT Part A	.43	CFT animal	.41	Digit span forward	.90
TMT Part B	.80	Boston naming test	.97		
ADAS cancellation	−.52	ADAS naming	.67		
Factor 4 (Language)		Factor 4 (Attention)		Factor 4 (Memory)	
CFT animal	.52	Digit span forward	.79	ADAS recognition	.47
Boston naming test	.56	Digit span backward	.68	RAVLT recall	.53
Factor 5 (Visuospatial processing)					
CDT command	.56				
CDT copy	.56				

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative. The factors are ordered according to sum squared loadings. Only variables with factor loadings >.40 are displayed in the table. HC = healthy control; MCI = mild cognitive impairment; AD = Alzheimer's disease; ADAS = Alzheimer's Disease Assessment Scale; RAVLT = Rey Auditory Verbal Learning test; CDT = Clock Drawing Test; CFT = Categorical Fluency Test; TMT = Trail Making Test.

recognition, RAVLT recall and RAVLT forgetting) or a subset of them always load on one factor; (b) instruments that are supposed to measure attention and working memory (digit span, forward and backward) always load on one factor with only one exception (AD, 24 months); (c) the two categorical fluency tests (CFT) and the two naming tests (Boston naming and ADAS naming) or a subset of them always load on one factor; (d) the three visuospatial processing tests (CDT command and copy, and ADAS construction) usually load on one factor, although ADAS construction was occasionally absent because its factor loading was not significant; and (e) the three executive function instruments (trail making test Parts A and B, and ADAS cancellation) or a subset of them load together either on the same factor with the three visuospatial processing tests or on one separate factor.

The second pattern is that across the four sessions: The EFAs point to a five-factor model (memory, visuospatial processing, attention, language, executive function) for the HC group, but a four-factor model (memory, visuospatial processing/executive function, attention, language) for the MCI (except for baseline) and AD groups. The five-factor model, particularly, is in strong agreement with prior theory-driven research on the ADNI battery (Johnson et al., 2012; Park et al., 2012). The difference between the four- and five-factor models is that the executive and visuospatial factors are distinct in the HC group but not in the MCI and AD groups. Although this suggests that the underlying factor structure may be different among the three diagnostic groups, that inference requires evidence from measurement invariance testing, which we now consider.

Single Group CFA Results

First, we created CFA models based on the EFA results and fitted them to the data of each diagnostic group in each session. Because the EFAs suggested that both a five-factor and a four-factor models might fit the data, we considered both. To preview, the five-factor model (Figure 1) delivered a better fit to the data. It turned out that although the four-factor model provided acceptable fits for the MCI group (CFI = .90, RMSEA = .06) and the AD group (CFI = .90, RMSEA = .07) at baseline, it provided a rather poor fit for the HC group (CFI = .83, RMSEA = .07). Consequently, we did not consider this model further. In contrast, the five-factor model delivered acceptable fits across the three diagnostic groups and across the four sessions.

By default, it is assumed that the underlying cognitive function accounts for all of the covariation between the tests loading on the given factor. Accordingly, there should be zero correlation between the residuals of the tests, which are the shared variance not explained

by the factor. However, in empirical research, this is often not true because tests that share similar assessment methods often have correlated residuals. In the current circumstance, we relaxed the residual correlations between ADAS naming test and Boston naming test based on the modification indices. Our addition of the particular residual correlation was justified on the ground that the two naming tests shared very similar testing procedures and response formats (Brown, 2015).

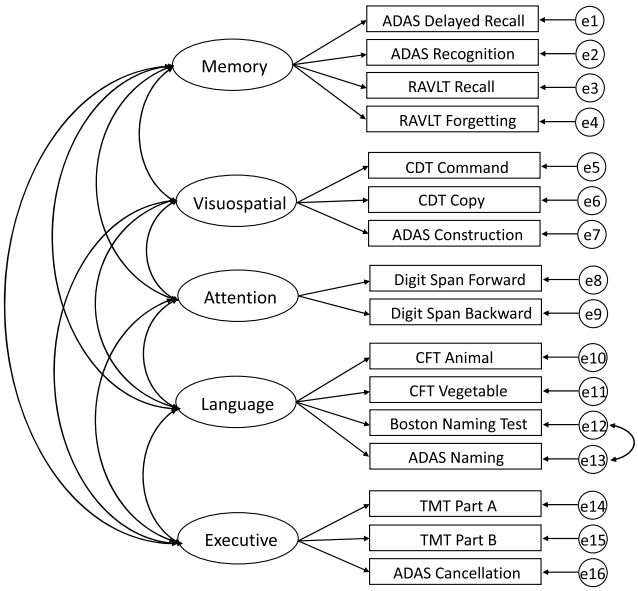
The final five-factor model is depicted in Figure 1, and its fit statistics and factor loadings are displayed in Table 4. The third and fourth rows in Table 4 show that CFIs ≥ .90 and RMSEAs ≤ .08 were nearly always achieved, establishing that the five-factor model yielded acceptable fits for all groups in all sessions. The only exception was RMSEA = .09 in the AD group at 24 months, where there was a Heywood case for the RAVLT recall variable (see the far-right column of Table 4). After examining the descriptive statistics, we concluded that the Heywood case was most likely due to range restriction in the RAVLT recall scores, which resulted from a floor effect in the AD group's recall at 24 months. Considering that the negative variance for RAVLT recall was very small (−.08), and the model converged, we continued to apply the five-factor model in the MG-CFAs.

MG-CFA Results

The MG-CFAs evaluated factorial invariance both across groups and across the 2-year testing interval. We separately report the fit statistics for factorial invariance between HC and AD groups (Table 5), between MCI and AD groups (Table 6), and between HC and AD groups (Table 7). In each pairwise comparison, the between-group invariance results are separately reported for each testing session (baseline, 6 months, 12 months, 24 months). The fit statistics for between-session factorial invariance within each diagnostic group are displayed in Table 8. In the following sections, we first discuss whether the factor structure remains stable between each pair of diagnostic groups and then consider whether the factor structure stays invariant over the 2-year span within each individual group.

Figure 1

The 5-Factor Model for the ADNI Alzheimer's Disease Neuroimaging Initiative Battery



Invariance Between HC and MCI

In Table 5, it can be seen that configural invariance, which implies the same number of factors and loading patterns across groups, was achieved between HC and MCI groups in all of the four testing sessions (CFIs ≥ .93 and RMSEAs ≤ .05). Thus, over the 2-year span, the HC and MCI groups' data were both satisfactorily captured by a five-factor model. Similarly, weak invariance was also consistently established between the HC and MCI groups (CFIs ≥ .91 and RMSEAs ≤ .06), and so was strong invariance (CFIs ≥ .90 and RMSEAs ≤ .06). These results show that the factor loadings and factor intercepts are equal between the two groups, over the 2-year interval. However, strict invariance was rejected in all of the sessions (CFIs ≤ .76 and RMSEAs ≥ .09), indicating that the residual variances were always unequal between HC and MCI subjects.

Invariance Between MCI and AD

The invariance results for MCI versus AD resemble those for HC versus MCI in the first three sessions but not in the last one. As can been seen in Table 6, configural invariance (CFIs ≥ .93 and RMSEAs ≤ .06), weak invariance (CFIs ≥ .91 and RMSEAs ≤ .07) and strong invariance (CFIs ≥ .91 and RMSEAs ≤ .07) were all achieved between the MCI and AD groups across the baseline, 6 months and 12 months sessions, while strict invariance was consistently rejected (CFIs ≤ .87 and RMSEAs ≥ .08). However, at 24 months, we were only able to establish configural invariance (CFIs = .92 and RMSEAs = .08) between the two groups. Modification indices suggest that the major reason for the misfits is constraining the factor loadings for TMT Part B (executive function), RAVLT recall and ADAS delayed recall (memory) to be equal across the two groups. The failure to establish weak invariance (CFI = .87 and RMSEAs = .09) means that strong and strict invariance cannot be achieved, because the model restrictions for those two levels of invariance incorporate those for weak invariance (Vandenberg & Lance, 2000; Wu et al., 2007).

To sum up, in the first year, MCI and AD subjects shared the same five-factor structure, the same factor loadings, and the same factor intercepts, although the residual variances differed between the two groups. However, this did not hold in the second year: The basic five-factor structure was still invariant between the two groups, but the factor loadings and intercepts were no longer equal.

Invariance Between HC and AD

The picture for the HC versus AD comparison is quite different from the two comparisons above. In Table 7, it can be seen that configural invariance was consistently obtained across the four sessions (CFIs ≥ .90 and RMSEAs ≤ .08). However, there was only partial support for weak invariance. Although weak invariance was established between HC and AD groups at 6 months and at 12 months (CFIs = .90 and RMSEAs = .06 and .07), it was not at

Table 4

Single Group Confirmatory Factor Analysis Results of the ADNI Neuropsychological Battery for the Three Diagnostic Groups Across the Four Sessions

	Baseline			6 months			12 months			24 months		
Indicators	HC	MCI	AD	HC	MCI	AD	HC	MCI	AD	HC	MCI	AD
Model fit statistics												
χ^2	147.48	185.19	160.85	99.38	167.52	169.08	138.52	174.05	141.82	135.92	127.44	275.58
Df	93	93	93	93	93	93	93	93	93	93	93	93
CFI	.91	.94	.92	.99	.95	.93	.93	.94	.96	.90	.96	.90
RMSEA	.05	.05	.06	.02	.05	.07	.05	.06	.05	.05	.05	.09
Factor loadings memory												
ADAS delayed recall	.65	.75	.48	.73	.81	.46	.71	.79	.35	.72	.82	.10
ADAS recognition	.28	.50	.43	.29	.50	.43	.35	.52	.47	.41	.49	.35
RAVLT recall	.82	.76	.95	.86	.73	.94	.91	.82	.83	.80	.86	1.04
RAVLT forgetting	−.59	−.62	−.28	−.57	−.68	−.42	−.44	−.64	−.21	−.45	−.60	−.12
Visuospatial												
CDT command	.83	.70	.77	.55	.84	.81	.87	.73	.76	.75	.80	.78
CDT copy	.57	.58	.70	.59	.57	.75	.65	.67	.73	.40	.56	.81
ADAS construction	.29	.48	.59	.29	.42	.59	.58	.56	.58	.12	.26	.77
Attention												
Digit span forward	.99	.55	.47	.81	.65	.56	.70	.52	.71	.63	.75	.66
Digit span backward	.60	.86	.82	.77	.84	.88	.87	.96	.71	.90	.75	.80
Language												
CFT Animal	.74	.75	.78	.75	.72	.77	.74	.72	.83	.71	.78	.89
CFT Vegetable	.57	.69	.78	.58	.71	.78	.59	.80	.79	.53	.76	.86
Boston naming test	.26	.36	.51	.19	.48	.49	.70	.59	.60	.12	.48	.74
ADAS naming	.41	.66	.61	.38	.65	.62	.49	.64	.68	.31	.61	.78
Executive												
TMT Part A	.59	.66	.74	.66	.75	.77	.77	.71	.82	.55	.74	.87
TMT Part B	.87	.84	.67	.89	.79	.68	.73	.91	.58	.71	.92	.63
ADAS cancellation	−.29	−.54	−.68	−.40	−.59	−.82	−.44	−.64	−.83	−.54	−.58	−.80

Note. HC = healthy control; MCI = mild cognitive impairment; AD = Alzheimer's disease; ADAS = Alzheimer's Disease Assessment Scale; RAVLT = Rey Auditory Verbal Learning Test; CDT = Clock Drawing Test; CFT = Categorical Fluency Test; TMT = Trail Making Test.

baseline or at 24 months (CFIs = .87 and .84, RMSEAs = .05 and .09). This shows that factor loadings were equal between the HC and AD groups in the 6- and 12-month sessions, but not in the other two sessions. Modification indices suggest that the major sources of the unsatisfactory fits were constraining the factor loadings for ADAS delayed recall (memory), CFT animal and vegetables (language), and TMT Part B (executive function) to be equal across the two groups. Because weak invariance was rejected for the data at

Table 5

Multigroup Confirmatory Factor Analysis Results of the ADNI Neuropsychological Battery Across the HC and MCI Groups Within the Four Sessions

Session	Model	χ2	df	CFI	RMSEA	Δχ2	Δdf	*P*	ΔCFI
Baseline									
	Configural	332.67	186	.93	.05	—	—	—	—
	Weak	381.62	197	.91	.06	48.94	11	<.01	.02
	Strong	423.46	208	.90	.06	41.85	11	<.01	.01
	Strict	726.13	224	.76	.09	302.67	16	<.01	.14
6 months									
	Configural	266.90	186	.96	.04	—	—	—	—
	Weak	344.41	197	.93	.05	77.51	11	<.01	.03
	Strong	351.69	208	.93	.05	7.28	11	.77	.00
	Strict	702.17	224	.76	.09	350.48	16	<.01	.17
12 months									
	Configural	312.57	186	.94	.05	—	—	—	—
	Weak	367.68	197	.92	.06	55.11	11	<.01	.02
	Strong	396.64	208	.91	.06	28.97	11	<.01	.01
	Strict	739.69	224	.75	.10	343.05	16	<.01	.16
24 months									
	Configural	263.36	186	.94	.05	—	—	—	—
	Weak	313.53	197	.91	.06	50.166	11	<.01	.03
	Strong	322.66	208	.91	.06	9.133	11	.61	.00
	Strict	618.62	224	.71	.10	295.955	16	<.01	.20

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative; HC = healthy control; MCI = mild cognitive impairment; RMSEA = root mean squared error; CFI = comparative fit index; df = degrees of freedom; HC = healthy control; MCI = mild cognitive impairment.

Table 6

Multigroup Confirmatory Factor Analysis Results of the ADNI Neuropsychological Battery Across the MCI and AD Groups Within the Four Sessions

Session	Model	χ2	df	CFI	RMSEA	Δχ2	Δdf	*p*	ΔCFI
Baseline									
	Configural	346.04	186	.93	.06	—	—	—	—
	Weak	404.67	197	.91	.06	58.63	11	<.01	.02
	Strong	430.22	208	.91	.06	25.55	11	<.01	.00
	Strict	602.98	224	.84	.08	172.77	16	<.01	.07
6 months									
	Configural	336.60	186	.94	.06	—	—	—	—
	Weak	415.37	197	.91	.07	78.78	11	<.01	.03
	Strong	427.32	208	.91	.06	11.94	11	.37	.00
	Strict	653.60	224	.83	.09	226.28	16	<.01	.08
12 months									
	Configural	315.88	186	.95	.05	—	—	—	—
	Weak	431.54	197	.91	.07	115.67	11	<.01	.04
	Strong	461.53	208	.91	.07	29.99	11	<.01	.00
	Strict	645.80	224	.84	.09	184.26	16	<.01	.07
24 months									
	Configural	403.02	186	.92	.08	—	—	—	—
	Weak	537.53	197	.87	.09	134.51	11	<.01	.05
	Strong	—	—	—	—	—	—	—	—
	Strict	—	—	—	—	—	—	—	—

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative; MCI = mild cognitive impairment; AD = Alzheimer's disease; RMSEA = root mean squared error; CFI = comparative fit index; df = degrees of freedom.

baseline and at 24 months, we only tested strong invariance for the 6- and 12-month sessions. For these two sessions, strong invariance was rejected (CFIs = .86 and .83, RMSEAs = .08 and .09), showing that factor intercepts are different for the HC and AD groups. Accordingly, strict invariance was not tested.

In summary, configural invariance was established between HC and AD subjects, indicating that a shared set of cognitive functions was being measured for these two groups. Weak invariance, however, only held at 6 months and at 12 months, meaning that factor loadings were sometimes not identical between HC and AD groups. Also, we found solid evidence against strong and strict invariance between the two groups, indicating that the factor intercepts and residual variances are always different between HC and AD groups.

Invariance Over Time

Now, we turn to factorial invariance across time within each diagnostic group. As shown in the third and fourth columns of

Table 7

Multigroup Confirmatory Factor Analysis Results of the ADNI Neuropsychological Battery Across the HC and AD Groups Within the Four Sessions

Session	Model	χ2	df	CFI	RMSEA	Δχ2	Δdf	*p*	ΔCFI
Baseline									
	Configural	308.33	186	.91	.06	—	—	—	—
	Weak	385.59	197	.87	.05	77.26	11	<.01	.04
	Strong	—	—	—	—	—	—	—	—
	Strict	—	—	—	—	—	—	—	—
6 months									
	Configural	268.461	186	.95	.05	—	—	—	—
	Weak	356.148	197	.90	.06	87.687	11	<.01	.05
	Strong	439.246	208	.86	.08	83.098	11	<.01	.04
	Strict	—	—	—	—	—	—	—	—
12 months									
	Configural	280.34	186	.95	.05	—	—	—	—
	Weak	379.05	197	.90	.07	98.71	11	<.01	.05
	Strong	519.54	208	.83	.09	140.48	11	<.01	.07
	Strict	—	—	—	—	—	—	—	—
24 months									
	Configural	411.502	186	.90	.08	—	—	—	—
	Weak	546.186	197	.84	.09	134.68	11	<.01	.06
	Strong	—	—	—	—	—	—	—	—
	Strict	—	—	—	—	—	—	—	—

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative; HC = healthy control; AD = Alzheimer's disease; RMSEA = root mean squared error; CFI = comparative fit index; df = degrees of freedom.

Table 8
Multigroup Confirmatory Factor Analysis Results of the ADNI Neuropsychological Battery Across the Four Sessions Within the Three Diagnostic Groups

Session	Model	χ^2	df	CFI	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI
HC									
	Configural	521.30	372	.93	.05	—	—	—	—
	Weak	586.55	405	.92	.05	65.25	33.00	<.01	.02
	Strong	663.17	438	.90	.05	76.61	33.00	<.01	.02
	Strict	790.61	486	.86	.06	127.44	48.00	<.01	.04
MCI									
	Configural	654.20	372	.95	.05	—	—	—	—
	Weak	696.63	405	.95	.05	42.43	33.00	.13	.00
	Strong	768.19	438	.94	.05	71.57	33.00	<.01	.01
	Strict	857.35	486	.93	.05	89.15	48.00	<.01	.01
AD									
	Configural	747.34	372	.92	.07	—	—	—	—
	Weak	813.83	405	.92	.07	66.50	33.00	<.01	.01
	Strong	863.03	438	.91	.07	49.20	33.00	.03	.00
	Strict	967.17	486	.90	.07	104.13	48.00	<.01	.01

Note. ADNI = Alzheimer's Disease Neuroimaging Initiative; RMSEA = root mean squared error; CFI = comparative fit index; df = degrees of freedom; HC = healthy control; MCI = mild cognitive impairment; AD = Alzheimer's disease.

Table 8, in all of the diagnostic groups, the CFIs and RMSEAs are within an acceptable range for configural invariance (CFIs ≥ .92, RMSEAs ≤ .07), weak invariance (CFIs ≥ .92, RMSEAs ≤ .07) and strong invariance (CFIs ≥ .90, RMSEAs ≤ .07). Therefore, the number of factors, factor loadings, and factor intercepts remained stable across the 2-year interval, and this is true in all of the three diagnostic groups. However, strict invariance was satisfied in the MCI (CFI = .93, RMSEA = .05) and AD groups (CFI = .90, RMSEA = .07), but not in the HC group (CFI = .86, RMSEA = .06). This means that while the residual variances are longitudinally invariant in the MCI and AD groups, they vary across sessions in the HC group.

Discussion

In this article, we used both exploratory (EFA) and confirmatory (CFA) factor analysis to evaluate factorial invariance for the ADNI neuropsychological battery, for three diagnostic groups (HC, MCI, AD) across 2 years of repeated assessments. With the exploratory analyses, the results converged on a five-factor model (memory, visuospatial processing, attention, language, executive function), which is consistent with an *a priori* conception that was proposed for the ADNI baseline data (Johnson et al., 2012; Park et al., 2012). With the confirmatory analyses, we established that after a minor modification, the five-factor model fit the data of each diagnostic group well within each session. When factorial invariance tests of this model were conducted, we found that although configural invariance always held across the three diagnostic groups, weak and strong invariance were only established between the HC and MCI groups for all four sessions and between the MCI and AD groups for the first three sessions. Weak invariance held only for certain sessions between the HC and AD groups, but strong invariance between the two groups was rejected in all sessions. Strict invariance was always rejected, across all groups in all sessions. In addition, we found that the factor structure remained stable across the 2-year testing interval; specifically, configural, weak, strong, and strict invariance were all satisfied, except for strict invariance in the HC group. A summary of the factorial invariance test results is presented in Table 9.

In the initial EFAs, the five-factor model always held for the HC data, whereas a four-factor model held in most sessions for the MCI and AD groups. The four-factor model differs from the five-model only in that it combines two of the factors of the latter model (visuospatial processing and executive functions). This combined factor is reasonable, considering the close associations that have been reported between these two cognitive domains (e.g., Libon et al., 1994; Miyake et al., 2001). After comparing the two models with single group CFAs, we concluded that the five-factor model yielded the best overall fit (see Table 4).

Next, the MG-CFAs showed that configural invariance held between the HC, MCI, and AD groups in each testing session, establishing that all three groups' neuropsychological performance was captured by the same five-factor structure. Therefore, we can conclude that the ADNI battery measures the same five cognitive abilities in all three diagnostic groups. Weak invariance was consistently established between the HC and MCI groups, but not always established between the MCI and AD groups or between the HC and AD groups. Statistically, weak invariance means that when test scores are regressed on their common factors, the regression slopes (factor loadings) are the same between the groups that are being compared. Because weak invariance is a precondition for comparing factor variances and covariances between groups, rejection of weak invariance threatens the soundness of between-group comparisons in correlation-based or criterion-based validity (Meredith & Teresi, 2006; Tuokko et al., 2009). The two most common examples are convergent and discriminant validity. In the current context, convergent validity measures whether neuropsychological instruments that are supposed to tap the same cognitive domain are in fact highly correlated, which is supported when the instruments all load highly on the corresponding factor. Discriminant validity measures whether instruments that are intended to measure distinct cognitive domains are not highly correlated, which is supported when the instruments load on different factors. As weak invariance was absent between the MCI and AD groups at 24 months, and between the HC and AD

Table 9
Summary of Multigroup Confirmatory Analysis Results

Types of invariance	Levels of invariance supported
Between-group invariance	
HC versus MCI – baseline	Configural, weak, strong
HC versus MCI – 6 months	Configural, weak, strong
HC versus MCI – 12 months	Configural, weak, strong
HC versus MCI – 24 months	Configural, weak, strong
MCI versus AD – baseline	Configural, weak, strong
MCI versus AD – 6 months	Configural, weak, strong
MCI versus AD – 12 months	Configural, weak, strong
MCI versus AD – 24 months	Configural
HC versus AD – baseline	Configural
HC versus AD – 6 months	Configural, weak
HC versus AD – 12 months	Configural, weak
HC versus AD – 24 months	Configural
Within-group longitudinal invariance	
HC	Configural, weak, strong
MCI	Configural, weak, strong, strict
AD	Configural, weak, strong, strict

Note. HC = healthy control; MCI = mild cognitive impairment; AD = Alzheimer's disease.

groups at baseline and 24 months, it is statistically inappropriate to compare convergent or discriminant validity between diagnostic groups in these particular sessions of the ADNI dataset.

In addition, although strong invariance was consistently observed between the HC and MCI groups and between the MCI and AD groups in three of the four sessions, it was never observed between the HC and AD groups. Statistically, the failure to establish strong invariance shows that the factor intercepts are not identical between groups (Beaujean, 2014; Brown, 2015). Consequently, the groups' average scores on particular instruments are not equal when factor scores are 0. Thus, even with the same factor-test regression slope across groups, the same test score will still be calibrated to different factor scores. Therefore, it is statistically inappropriate to compare factor scores between MCI and AD groups during one of the sessions (24 months) or between HC and AD groups in any session. This is an important limitation considering that some researchers have argued for the discriminant or predictive power of factor scores (Chapman et al., 2010; Giraldo et al., 2017). Apart from factor scores, another common practice for comparing group means is to calculate group differences in composite scores. However, it is often overlooked that it is only justifiable to compare composite scores when observed means are equally calibrated to latent means across groups, which requires equal factor loadings and intercepts (Steinmetz, 2013). Therefore, if strong invariance is rejected between HC and AD groups and sometimes between MCI and AD groups, this may threaten the legitimacy of findings based on between-group comparisons in composite scores for the ADNI battery.

The picture for between-session factorial invariance is quite different. In all three groups, the five-factor structure was stable across the four sessions, satisfying all four levels of invariance, with only a minor departure from the strict invariance criterion in the HC group. Thus, within each diagnostic group, the factor configuration, factor loadings, and factor intercepts were all longitudinally invariant, and thus, convergent and divergent validity, test scores, and the means of factor scores can all be directly compared between sessions. This is especially meaningful when it comes to tracking longitudinal changes in cognitive functions. For instance, in longitudinal research, the latent growth model (LGM) is a commonly used statistical method that measures both group-level growth and individual differences in growth. However, the use of this method is only warranted when weak and strong invariance are established longitudinally (Ferrer et al., 2008; Vandenberg & Lance, 2000). Because we found that both weak and strong invariance held across a 2-year span within each ADNI diagnostic group, LGM would be an ideal tool to analyze longitudinal patterns of the ADNI data within single diagnostic groups.

Overall, the EFA and MG-CFA results agree on two key points: Factor structures are invariant across testing sessions but not across diagnostic groups. There is a slight discrepancy between the EFAs and CFAs with respect to the number of factors per diagnostic group. EFAs suggest a five-factor structure for the HC group and a four-factor structure for the MCI and AD groups, but the MG-CFAs showed that a five-factor structure held across the three diagnostic groups (configural invariance). Such a discrepancy is not surprising, considering that the number of factors is specified *a priori* in CFA, but it is based on both statistical criteria (eigenvalues, scree plots, and parallel analysis) and theoretical criteria (e.g., parsimony, interpretation) in EFA.

It is worth mentioning that the current findings are broadly consistent with Park et al. (2012). Those authors analyzed the baseline ADNI data and established all four levels of invariance between two subject groups: less versus more functionally impaired subjects. Here, the points of disagreement between our results and Park et al.'s are most likely due to differences in the subject groups that were compared. Our subject groups reflected clinical diagnoses, whereas Park et al. performed a median split on the CDR-SOB (sum of scores) and divided the total ADNI sample into less and more impaired groups. They argued that this avoided circularity that would be inherent in using the ADNI neuropsychological battery to form diagnostic groups as well as to conduct factor analyses. However, we thought it was advisable to use the clinical diagnoses for group classification for three reasons.

First, to our minds, there is no circularity problem because the battery test scores are not the only basis for ADNI diagnoses. Clinicians based those diagnoses on multiple sources of information, which included medical histories, genetic data, plasma and serum biomarkers, MRI data, and neuropsychological battery test scores. However, only the battery test scores figured in our factor analyses. Second, if the use of the clinical diagnoses to form comparison groups creates a circularity problem, it is unclear how using CDR scores to form comparison groups avoids that problem. That is because different CDR cutoffs were used for different diagnostic group in the ADNI screening procedure (see the Method section), making CDR scores one of the determinants of diagnostic group membership. From a psychometric point of view, relying on CDR scores would simply be a less reliable method of forming comparison groups than relying on the diagnoses. Third, the median split method has often been criticized for reducing statistical power and increasing Type II error (e.g., McClelland et al., 2015; Rucker et al., 2015), which we hope to avoid. More importantly, median splits will inevitably increase within-group heterogeneity by mixing subjects who differ in severity of cognitive impairment. Thus, establishing factorial invariance between groups that have been formed using median splits, even if psychometrically acceptable, is less clinically

meaningful. This is a rather important consideration, as clinical application is a core motivation of ADNI research (e.g., Johnson et al., 2012).

Last, we should acknowledge that our sample size is smaller than some of the prior studies that examined factorial invariance in aging and dementia (e.g., Hayden et al., 2011; Siedlecki et al., 2008). Currently, there are still debates about what sample size is ideal for factor analyses (Kline, 2005; Maccallum et al., 1999; Wolf et al., 2013). In empirical research, the most frequently quoted rule of thumb is a minimum sample size of 100 or 200 (Boomsma, 1982, 1985) and 10 observations per variable (Nunnally, 1994). Our sample satisfies such requirements. Meanwhile, we only included variables with KMO > .5 in our analyses, which ensures sampling adequacy. For these reasons, sample size should not be a threat to the validity of our results.

Conclusion

For the three ADNI diagnostic groups, we established configural invariance and found some support for weak invaraince across these groups. This suggests that the tests in the ADNI neuropsychological battery tap the same cognitive functions in all of these groups. At a more fine-grained level, however, the failure to establish strong and strict invariance argues for caution in making between-group inferences about cognitive functions based on group differences in test scores or factor scores. Without these forms of invariance, the quantitative relation between scores on the individual tests and true scores of the cognitive functions that they measure may not be the same for different diagnostic groups. In addition, we established configural, weak, and strong invariance across the 2 year testing interval, for each diagnostic group. This is particularly important when it comes to interpreting longitudinal changes in test scores within each diagnostic group. It shows that between-session declines in test scores are due to changes in the underlying cognitive functions, rather than to a change in what the tests measure.

References

Alzheimer's Association. (2020). 2020 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *16*(3), 391–460. https://doi.org/10.1002/alz.12068

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*(2), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction: Vol. Part I* (pp. 149–173). North-Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. https://doi.org/10.1007/BF02294248

Brainerd, C. J., Reyna, V. F., Gomes, C. F. A., Kenney, A. E., Gross, C. J., Taub, E. S., Spreng, R. N., & Alzheimer's Disease Neuroimaging Initiative. (2014). Dual-retrieval models and neurocognitive impairment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 41–65. https://doi.org/10.1037/a0034057

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Press.

Browne, M., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing Structural Equation Models* (pp. 136–162). Sage Publications.

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*(3), 289–325. https://doi.org/10.1207/s15327906mbr1203_2

Chapman, R. M., Mapstone, M., Porsteinsson, A. P., Gardner, M. N., McCrary, J. W., DeGrush, E., Reilly, L. A., Sandoval, T. C., & Guillily, M. D. (2010). Diagnosis of Alzheimer's Disease using neuropsychological testing improved by multivariate analyses. *Journal of Clinical and Experimental Neuropsychology*, *32*(8), 793–808. https://doi.org/10.1080/13803390903540315

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Exploratory Factor Analysis*, *10*(7), 1–9. https://doi.org/10.7275/jyj1-4868

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., Jones, R. N., Mukherjee, S., Curtis, M. S., Harvey, D., Weiner, M., & Mungas, D. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, *6*(4), 502–516. https://doi.org/10.1007/s11682-012-9186-z

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100: Historical developments and future directions*. Erlbaum.

de Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the Memory Compensation Questionnaire. *Psychological Assessment*, *17*(2), 168–178. https://doi.org/10.1037/1040-3590.17.2.168

Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M., & Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment. *Journal of the International Neuropsychological Society*, *9*(6), 936–946. https://doi.org/10.1017/S1355617703960139

Dowling, N. M., Hermann, B., La Rue, A., & Sager, M. A. (2010). Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. *Neuropsychology*, *24*(6), 742–756. https://doi.org/10.1037/a0020176

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(1), 22–36. https://doi.org/10.1027/1614-2241.4.1.22

Flores, I., Casaletto, K. B., Marquine, M. J., Umlauf, A., Moore, D. J., Mungas, D., Gershon, R. C., Beaumont, J. L., & Heaton, R. K. (2017). Performance of Hispanics and Non-Hispanic Whites on the NIH Toolbox Cognition Battery: The roles of ethnicity and language backgrounds. *The Clinical Neuropsychologist*, *31*(4), 783–797. https://doi.org/10.1080/13854046.2016.1276216

Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., Curtis, S. M., Mungas, D., & Crane, P. K. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior*, *6*(4), 517–527. https://doi.org/10.1007/s11682-012-9176-1

Giraldo, D. L., Sijbers, J., & Romero, E. (2017). Quantifying cognition and behavior in normal aging, mild cognitive impairment, and Alzheimer's disease. *13th International Conference on Medical Information Processing and Analysis*, *10572*, Article 105720H. https://doi.org/10.1117/12.2287036

Goodglass, H., Kaplan, E., & Weintraub, S. (1983). *Boston naming test*. Lea & Febiger.

Gustafsson, J. E., & Stahl, P. A. (2005). *STREAMS 3.0 User's Guide*. MultivariateWare.

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

Hayden, K. M., Jones, R. N., Zimmer, C., Plassman, B. L., Browndyke, J. N., Pieper, C., Warren, L. H., & Welsh-Bohmer, K. A. (2011). Factor structure of the National Alzheimer's Coordinating Centers Uniform Dataset Neuropsychological Battery: An evaluation of invariance between and within groups over time. *Alzheimer Disease and Associated Disorders*, *25*(2), 128–137. https://doi.org/10.1097/WAD.0b013e3181ffa76d

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *134*(824), 486–501. https://doi.org/10.1098/rspa.1931.0209

Hirschfeld, G., & von Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R—A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, *19*(7), 1–11. https://doi.org/10.7275/qazy-2946

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

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), 1–55. https://doi.org/10.1080/10705519909540118

Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*(1), 90–98. https://doi.org/10.1016/j.jcps.2009.09.003

Johnson, D. K., Storandt, M., Morris, J. C., Langford, Z. D., & Galvin, J. E. (2008). Cognitive profiles in dementia: Alzheimer disease versus healthy brain aging. *Neurology*, *71*(22), 1783–1789. https://doi.org/10.1212/01.wnl.0000335972.35970.70

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

Jones, S. N., & Ayers, C. R. (2006). Psychometric properties and factor structure of an expanded CERAD neuropsychological battery in an elderly VA sample. *Archives of Clinical Neuropsychology*, *21*(4), 359–365. https://doi.org/10.1016/j.acn.2006.03.004

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/BF02289233

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117.

Kanne, S. M., Balota, D. A., Storandt, M., McKeel, D. W., & Morris, J. C. (1998). Relating anatomy to function in Alzheimer's disease: Neuropsychological profiles predict regional neuropathology 5 years later. *Neurology*, *50*(4), 979–985. https://doi.org/10.1212/WNL.50.4.979

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage Publications.

Libon, D. J., Glosser, G., Malamut, B. L., Kaplan, E., Goldberg, E., Swenson, R., & Prouty Sands, L. (1994). Age, executive functions, and visuospatial functioning in healthy older adults. *Neuropsychology*, *8*(1), 38–43. https://doi.org/10.1037/0894-4105.8.1.38

Lindeman, R. H. (1980). *Introduction to bivariate and multivariate analysis*. Scott Foresman.

Loehlin, J. C., & Beaujean, A. A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Taylor & Francis.

MacCallum, R. C., Widaman, K. F., Zhang, S., Hong, S., Maccallum, R. C., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

McClelland, G. H., Lynch, J. G., Jr., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false–positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology*, *25*(4), 679–689. https://doi.org/10.1016/j.jcps.2015.05.006

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *[JSTOR.]. Medical Care*, *44*(11), S69–S77.

Mitchell, M. B., Shaughnessy, L. W., Shirk, S. D., Yang, F. M., & Atri, A. (2012). Neuropsychological test performance and cognitive reserve in healthy aging and the Alzheimer's Disease spectrum: A theoretically-driven factor analysis. *Journal of the International Neuropsychological Society*, *18*(6), 1071–1080. https://doi.org/10.1017/S1355617712000859

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, *43*, 2412–2414. https://doi.org/10.1212/WNL.43.11.2412-a

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005a). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, *15*(4), 869–877. https://doi.org/10.1016/j.nic.2005.09.008

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005b). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, *1*(1), 55–66. https://doi.org/10.1016/j.jalz.2005.06.003

Mungas, D., Widaman, K. F., Reed, B. R., & Tomaszewski Farias, S. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology*, *25*(2), 260–269. https://doi.org/10.1037/a0021090

Nunnally, J. C. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Park, L. Q., Gross, A. L., McLaren, D. G., Pa, J., Johnson, J. K., Mitchell, M., Manly, J. J., & Alzheimer's Disease Neuroimaging Initiative. (2012). Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging and Behavior*, *6*(4), 528–539. https://doi.org/10.1007/s11682-012-9190-3

Pedraza, O., Lucas, J. A., Smith, G. E., Willis, F. B., Graff-Radford, N. R., Ferman, T. J., Petersen, R. C., Bowers, D., & Ivnik, R. J. (2005). Mayo's older African American normative studies: Confirmatory factor analysis of a core battery. *Journal of the International Neuropsychological Society*, *11*(2), 184–191. https://doi.org/10.1017/S1355617705050204

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, *256*(3), 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

Petersen, R. C. (2011). Mild Cognitive Impairment. *New England Journal of Medicine*, *364*(23), 2227–2234. https://doi.org/10.1056/NEJMcp0910237

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Rawlings, A. M., Bandeen-Roche, K., Gross, A. L., Gottesman, R. F., Coker, L. H., Penman, A. D., Sharrett, A. R., & Mosley, T. H. (2016). Factor structure of the ARIC-NCS Neuropsychological Battery: An evaluation of invariance across vascular factors and demographic characteristics. *Psychological Assessment*, *28*(12), 1674–1683. https://doi.org/10.1037/pas0000293

Reitan, R. M., & Wolfson, D. (1985). The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation (Vol. *4*). *Reitan Neuropsychology*.

Revelle, W. (2016). *How to: Use the psych package for factor analysis and data reduction*. https://rdrr.io/cran/psychTools/f/inst/doc/factor.pdf

Rey, A. L. (1964). *Presses Universitaires de France*. Presses Universitaires de France.

Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry*, *141*(11), 1356–1364. https://doi.org/10.1176/ajp.141.11.1356

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, *25*(4), 666–678. https://doi.org/10.1016/j.jcps.2015.04.004

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Siedlecki, K. L., Honig, L. S., & Stern, Y. (2008). Exploring the structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. *Neuropsychology*, *22*(3), 400–411. https://doi.org/10.1037/0894-4105.22.3.400

Siedlecki, K. L., Manly, J. J., Brickman, A. M., Schupf, N., Tang, M.-X., & Stern, Y. (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology*, *24*(3), 402–411. https://doi.org/10.1037/a0017515

Steiger, J. (1990). Structural model evaluation and modification—an interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

Tuokko, H. A., Chou, P. H. B., Bowden, S. C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *Journal of the International Neuropsychological Society*, *15*(3), 416–425. https://doi.org/10.1017/S1355617709090602

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*(1), 4–70. https://doi.org/10.1177/109442810031002

Wechsler, D. (1987). *Wechsler memory scale-revised*. Psychological Corporation.

Weiner, M. W., & Veitch, D. P. (2015). Introduction to special issue: Overview of Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *11*(7), 730–733. https://doi.org/10.1016/j.jalz.2015.05.007

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Luthman, J., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L., Shen, L., Schwarz, A., Toga, A. W., & Trojanowski, J. Q. (2015). 2014 Update of the Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, *11*(6), e1–e120. https://doi.org/10.1016/j.jalz.2014.11.001

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L. M., Toga, A. W., & Trojanowski, J. Q. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, *13*(4), e1–e85. https://doi.org/10.1016/j.jalz.2016.11.007

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3), 1–13. https://doi.org/10.33151/ajp.8.3.93

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wu, A., Li, Z., & Zumbo, B. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1–26. https://doi.org/10.7275/mhqa-cd89

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94.

Yuan, K.-H., & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical & Statistical Psychology*, *54*(1), 161–175. https://doi.org/10.1348/000711001159366

Received September 22, 2020
Revision received December 26, 2020
Accepted February 10, 2021 ■