Article

Questioning the Meaning of a Change on the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog): Noncomparable Scores and Item-Specific Effects Over Time

Hugo Cogo-Moreira[1,2]* (iD), **Saffire H. Krance[3,4]*** (iD), **Sandra E. Black[3,4],**
Nathan Herrmann[3,4], Krista L. Lanctôt[3,4], Bradley J. MacIntosh[3,4], Michael Eid[1†],
and Walter Swardfager[3,4†], for the Alzheimer's Disease Neuroimaging Initiative[††]

Abstract

Longitudinal invariance indicates that a construct is measured over time in the same way, and this fundamental scale property is a *sine qua non* to track change over time using ordinary mean comparisons. The Alzheimer's Disease Assessment Scale–cognitive (ADAS-Cog) and its subscale scores are often used to monitor the progression of Alzheimer's disease, but longitudinal invariance has not been formally evaluated. A configural invariance model was used to evaluate ADAS-Cog data as a three correlated factors structure for two visits over 6 months, and four visits over 2 years (baseline, 6, 12, and 24 months) among 341 participants with Alzheimer's disease. We also attempted to model ADAS-Cog subscales individually, and furthermore added item-specific latent variables. Neither the three-correlated factors ADAS-Cog model, nor its subscales viewed unidimensionally, achieved longitudinal configural invariance under a traditional modeling approach. No subscale achieved scalar invariance when considered unidimensional across 6 months or 2 years of assessment. In models accounting for item-specific effects, configural and metric invariance were achieved for language and memory subscales. Although some of the ADAS-Cog individual items were reliable, comparisons of summed ADAS-Cog scores and subscale scores over time may not be meaningful due to a lack of longitudinal invariance.

Keywords

Alzheimer's disease, cognition, longitudinal invariance, reliability, structural equation modeling

In longitudinal studies, ordinary means of scales derived from the aggregation of individual items are often compared in order to answer questions like "How does a group progress over time?" or "Is a drug having a beneficial effect?". To answer these questions, statistical techniques such as paired *t* tests or repeated measures analyses of variance require that the continuous outcome is evaluating the underlying phenomenon in the same way at each repeated assessment. When an instrument and its items behave differently between different occasions of measurement, the scores cannot be meaningfully compared. The use of the same scale does not guarantee that psychometric features are being captured in the same way over time. For example, if levels of performance change sufficiently to cause ceiling or floor effects, an instrument of measurement must change to enable proper estimation of that change (Embretson, 2006). Assessing evidence of longitudinal invariance permits eval-

[1]Freie Universität Berlin, Berlin, Germany
[2]Universidade Federal de São Paulo, São Paulo, Brazil
[3]Sunnybrook Research Institute, Toronto, Ontario, Canada
[4]University of Toronto, Toronto, Ontario, Canada

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

Corresponding Author:
Walter Swardfager, Department of Pharmacology and Toxicology, University of Toronto, 1 King's College Circle, Toronto, Ontario, Canada M5S1A8.
Email: w.swardfager@utoronto.ca

uation of whether a scale assesses a given construct uniformly, on the same metric, over time.

Longitudinal invariance testing is conducted within the context of structural equation modeling, using item parameters such as factor loadings and thresholds (when the items are categorical) or intercepts (when items are continuous). The first level of invariance is called *configural invariance* and in longitudinal models it is achieved when the relationship between the latent variables (i.e., factors/domains of a given scale) and their manifest indicators are uniform across occasions; in other words, the number of factors and the general loading of the items onto those factors do not change from visit to visit. In the case of the ADAS-Cog, we would expect its 11 items to load equivalently onto the three subdomains (memory, praxis, and language) over time. A *sine qua non* to evaluate changes in means of aggregated scores in an ordinary paired *t* test or a repeated measures analysis of variance (ANOVA) is to guarantee that some level of invariance is observed across these parameters. Operationally, a series of constraints on factor loadings, and then on thresholds/intercepts, is imposed in a hierarchical order, and changes to the goodness of fit of the measurement model are evaluated (Van de Schoot et al., 2012; Van de Schoot et al., 2015).

The ADAS-Cog is commonly used to monitor the progression of cognitive symptoms in dementia due to Alzheimer's Disease (AD), and it has been used as a primary cognitive outcome measure across many large-scale randomized clinical trials (Connor & Sabbagh, 2008; Honig et al., 2018; D.-D. Li et al., 2019; Salloway et al., 2014; Weyer et al., 1997). It is also employed in cohort studies to track AD trajectories as, for example, one of many cognitive assessments performed in the Alzheimer's Disease Neuroimaging Initiative (ADNI; Weiner et al., 2013), the Japanese ADNI (Yagi et al., 2019), and others. The most commonly used version of this test is the 11-item ADAS-Cog, composed of three subdomains intended to assess learning and memory, language production and comprehension, and praxis (Rosen et al., 1984). Given that ADAS-Cog scores have been used as a clinical a trial endpoint to assess the efficacy of drug therapies, it is of importance to understand how confident one can be that repeated measurements capture cognition in the same way.

Some psychometric features of the ADAS-Cog have been described previously, and measurement issues have motivated numerous attempts to improve the ADAS-Cog. A recent review noted that 31 modified versions appear in the literature (Kueper et al., 2018), including an empirical solution presented by Verma et al. (2015), based on item response theory. Cano et al. (2010), Hobart et al. (2013), and Karin et al. (2014) identified ceiling effects, even among AD patients of mild to moderate severity. Longitudinally, some studies have found low reliability for measuring change (Grochowalski et al., 2016) and low

test–retest reliability for several of the individual items (Karin et al., 2014). In terms of convergent validity with clinical assessments over time, improvement on the ADAS-Cog has been related to clinical improvement, but many people who declined on the ADAS-Cog did not show clinical decline (Rockwood et al., 2007). Despite these observations, the ADAS-Cog and its three subdomains have been evaluated scarcely for measurement invariance. As part of a larger study, Dowling et al. (2016) conducted longitudinal invariance testing using a traditional modeling approach; however, their model specifications, and methods for treating different natures of items (i.e., categorical and continuous) were not described. Beyond that, they noted that some items exhibited different sensitivities to between-person differences at baseline versus changes over time, which hinted at the possible need for a more flexible modeling approach.

The present study aims to examine longitudinal measurement invariance of the ADAS-Cog and its subdomains among patients with mild AD using two approaches: a traditional approach related to the works of Millsap (2012) and Meredith (1993), and an alternative more flexible approach that includes item-specific effects, reducing the degrees of freedom (Eid et al., 2016; Eid & Kutscher, 2014). The latter approach relaxes the assumption that the relationships between the observed indicators and the underlying constructs are consistent between each occasion of measurement. We provide an empirical examination in the context of the ADNI, a naturalistic longitudinal study, making use of the 11 indicators of the ADAS-Cog across four visits: baseline, 6, 12, and 24 months. A better understanding of these fundamental scale properties might offer some guidance in the use of the ADAS-Cog items and their summed scores in longitudinal studies, including observational studies and randomized clinical trials.

Method

Sample

Data were obtained from the ADNI database (http://adni.loni.usc.edu/) in January, 2018. The ADNI was launched in 2003, as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD, with the primary goal of determining whether neuroimaging, other biomarkers, and clinical and neuropsychological assessments could be combined to measure the progression of early AD. For up-to-date information, see www.adni-info.org. Briefly, ADNI recruited participants between the ages of 55 and 90 years in North America to be followed longitudinally, with testing at regular intervals. A diagnosis of mild AD was determined using National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association criteria, as well a score of between

Table 1. Demographic and Dementia Characteristics (N = 341).

Characteristic	Baseline	6 Months	12 Months	24 Months
Baseline age, M (SD)	75 (5)			
Years of education, M (SD)	15 (3)			
ApoE ε4 carriers (% carriers)[a]	66			
Sex (% male)	55			
MMSE, M (SD)	23.2 (2.1)	22.2 (3.7)	20.9 (4.5)	18.7 (5.7)
CDR, M (SD)	0.76 (0.26)	0.89 (0.38)	1.04 (0.54)	1.27 (0.66)

Note. ApoE = apolipoprotein E; MMSE = Mini Mental Status Examination (score from 30 to 0; higher scores are better); CDR = Clinical Dementia Rating (rating from 0 to 3; 0 = no dementia, 0.5 = very mild dementia, 1.0 = mild dementia, 2.0 = moderate dementia, etc.).
[a]Data were not available for two AD.

20 and 26 on the Mini Mental State Examination (MMSE; a tool frequently used to screen for cognitive impairment in clinical, research, and community settings; (Arevalo-Rodriguez et al., 2015) and a score of 0.5 or 1.0 on the Clinical Dementia Rating Scale (CDR), indicating a very mild, or mild clinical level of impairment, respectively. At the time of data download, the ADNI database included 341 participants with AD at baseline, all of whom were included in the current study; their baseline MMSE scores, years of education, age, sex, and carrier status for the apolipoprotein E ε4 allele (ApoE ε4), a genetic variant associated with the risk of late-onset AD (Saunders et al., 1993), can be found in Table 1.
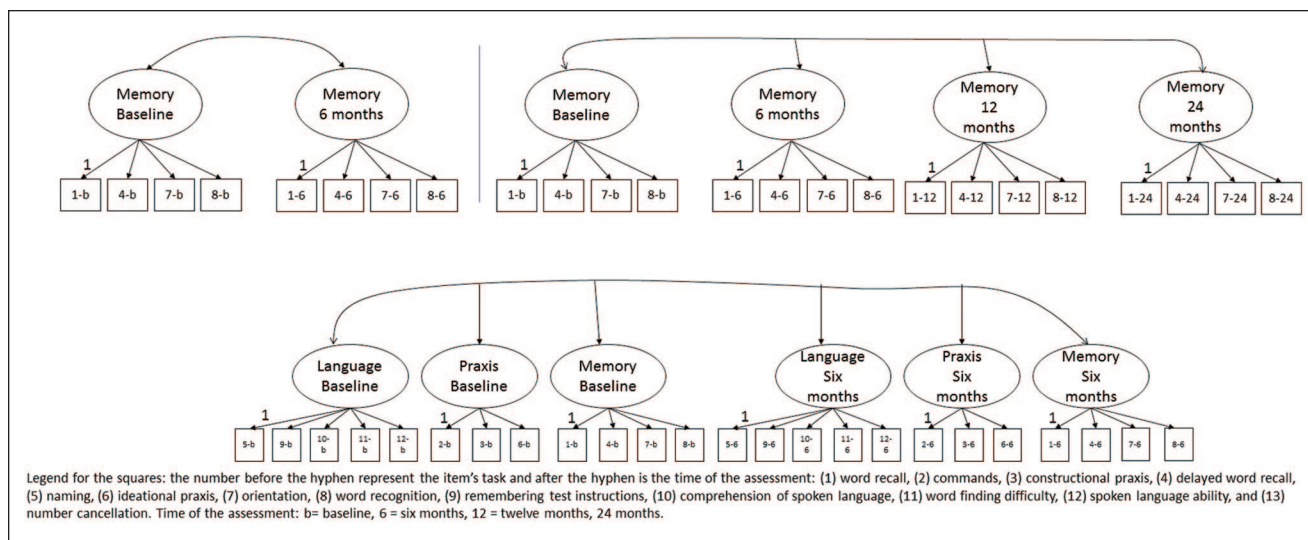
The ADAS-Cog

The ADAS-Cog was conducted by an Alzheimer's Disease Cooperative Study—ADAS certified psychometrist on ADNI participants at their baseline, 6-month, and subsequent annual visits. A detailed description of test administration is provided in the ADNI procedures manual (http://adni.loni.usc.edu/). The 13-item version of the ADAS-Cog was conducted (ADAS-Cog 13), allowing for calculation of either the 13-item total score (out of 85), or the more commonly used 11-item total score (out of 70), where a higher score indicated poorer performance and greater impairment. Test items consisted of the following tasks: (1) word recall, (2) commands, (3) constructional praxis, (4) delayed word recall, (5) naming, (6) ideational praxis, (7) orientation, (8) word recognition, (9) remembering test instructions, (10) comprehension of spoken language, (11) word finding difficulty, (12) spoken language ability, and (13) number cancellation. An alternate list of words was used for the Word Recall Task (Item 1) at Month 6, but the original list was used at Baseline, Month 12, and Month 24, which were considered sufficiently distanced in time to avoid practice effects. This also affects the Delayed Word Recall Task (Item 4), which asks the participant to recall words from the list (Item 1).

Statistical Analysis

For the ADAS-Cog 11 structure, and for the empirical factor structure later proposed by Verma et al. (2015), traditional invariance testing was conducted as described by Horn and McArdle (1992; Meredith & Horn, 2001), for each subdomain separately (Figure 1, top depicts, e.g., the *memory* subdomain as an unidimensional solution, where all the items related to this subdomain are loaded onto a single factor; the same procedure was conducted for Praxis and Language) and also for the three-correlated factor solution (Figure 1, bottom). For the unidimensional solution, longitudinal invariance was tested considering (1) a 6 months gap between the assessment (i.e., pre–post evaluation; two visits) and (2) across 2 years (i.e., four visits; baseline, 6 months, 12 months, and 24 months after baseline). For the three-correlated factors solution, evaluation across four visits was not conducted due to failure of invariance testing over the initial two visits. Only the Verma et al. (2015) and the ADAS-Cog 11 solutions could be tested under three correlated factors; the ADAS-Cog 13 structure is not admissible because at least two items per factor are required.

Subdomains were investigated separately under unidimensional (single-factor) solutions, specified as per guidance from previous literature. For the memory subdomain, three versions were analyzed; first, as per the ADAS-Cog 11, the memory score was based on four items (i.e., word recall, orientation, remembering test instructions, and word recognition); second, as per the ADAS-Cog 13, the memory score was based on five items (i.e., as per the ADAS-Cog 11 plus Item 4 [delayed word recall]); and third, constituted by four items, as suggested by Verma et al. (2015) to have longitudinal validity (i.e., word recall, orientation, word recognition, and delayed word recall; Mohs et al., 1997; Verma et al., 2015). For the language subdomain, two versions were analyzed; first, as per the ADAS-Cog 11 and 13, this subdomain was constituted by five items (i.e., commands, naming, comprehension of spoken language, word finding difficulty, and spoken language ability); second, as per

Figure 1. Traditional invariance testing of the three ADAS-Cog traits proposed by Verma et al., showing memory as a unidimensional construct across two and four visits (Figure 1, top, left, and right respectively) and the three-correlated factor structure (Figure 1, bottom) across two visits. Note that for language and praxis as unidimensional models (note shown here), the model specification would follow the structure shown on top (left and right side).
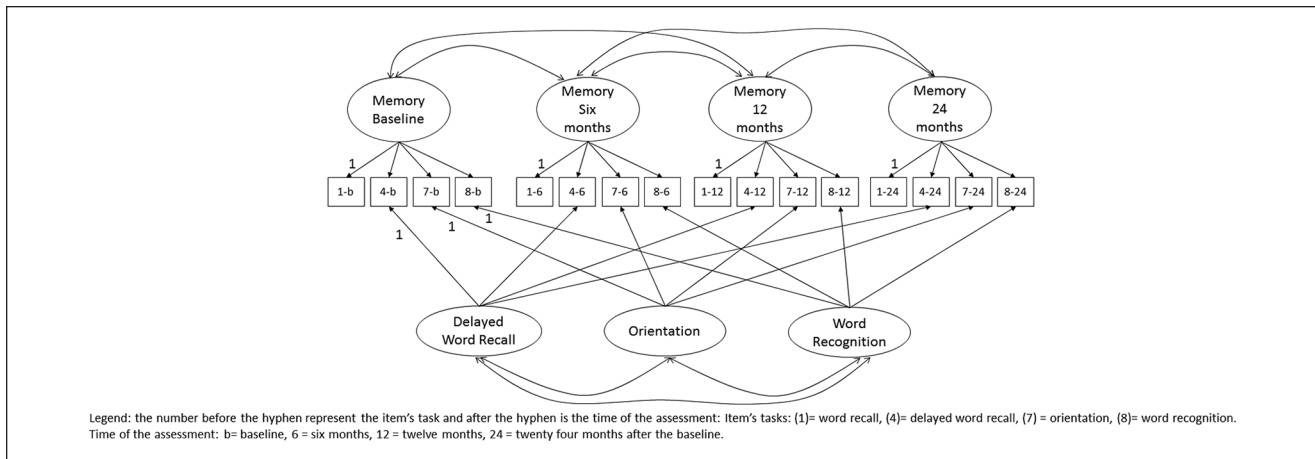
Verma et al. (2015) this subdomain was constituted by five items (i.e., naming, remembering test instructions, comprehension of spoken language, word finding difficulty, and spoken language ability). For praxis, only the solution of Verma et al. (2015) could be properly evaluated, which was constituted by three items (i.e., commands, constructional praxis, and ideational praxis); in a cross-sectional design fewer than three items per factor would result in a *just-specified* model for which it is not possible to evaluate model fit indices (Bollen, 1998).

The less restrictive level of invariance, the configural model, holds the same pattern of fixed and free factor loadings across time. In that model, thresholds (for the categorical items [language and praxis domains]) and intercepts (for the continuous items [for the memory domain]) and residual variances of the items were freely estimated across time. If configural invariance was achieved, other constraints were added progressively. Testing metric invariance, or "weak invariance," is said to be achieved if the association of a given item and its underlying factor is equal across the time. It is tested by imposing factor loadings to be equal across time as a further model constraint. Further restrictions impose constraints to thresholds/intercepts (called "scalar invariance" or "strong invariance") and, then, to residual variance (called "strict invariance"), and, last, constraints to the factor level variance.

For the models with four visits over time, we additionally took into to account item-specific effects (Eid et al., 2016; Eid & Kutscher, 2014), which gives more flexibility by reducing the degrees of freedom compared with traditional invariance testing. With the addition of latent variables capturing item-specific effects, the model becomes less restrictive, modeling at the same time the occasion effect, sometimes called state variance (Eid & Kutscher, 2014), and the item-specific effects (variance that is uniform across time), and as consequence it is possible to disentangle both of these sources of information. Figure 2 shows configural invariance testing for the memory subscale (Items 1, 4, 7, and 8) including item-specific effects across four visits. For example, the *orientation* items across the four visits were loaded onto an extra latent variable called *orientation*, and the items for *word recognition* subscale were loaded onto an extra latent factor called *word recognition*; for each item-specific factor, one loading parameter was fixed to 1 and the other factor loadings were free to be estimated, and means were fixed at zero. The items for *word recall* were used as a reference methodological factor by not loading the word recall items to any latent factor. Therefore, the number of item-specific factors to be specified will be the number of different subscale items less one, which is sufficient to model item-specific effects under configural invariance testing (Eid, 1996; Geiser & Lockhart, 2012). Together with the item-specific factor, the occasion factors, as previously specified (i.e., items answered at the same visit), and their hierarchical invariance restrictions (configural, weak, and scalar) were estimated together with the item-specific factor. The occasion and item-specific factors are orthogonal to one another, but the correlations of the indicator-specific factors are freely estimated in the same way that the occasion factor is allowed to be correlated and freely estimated.

Legend: the number before the hyphen represent the item's task and after the hyphen is the time of the assessment: Item's tasks: (1)= word recall, (4)= delayed word recall, (7) = orientation, (8)= word recognition.
Time of the assessment: b= baseline, 6 = six months, 12 = twelve months, 24 = twenty four months after the baseline.

Figure 2. Invariance testing under the specification of Eid et al. (2014; Eid et al., 2016) allowing configural invariance testing for the memory subscale proposed by Verma et al. (Items 1, 4, 7, and 8) as an example. Item-specific effects across four visits were added. Note that the same modeling approach was conducted for the other subscales (i.e., praxis and language).

Model fit was evaluated based on the following fit measures and their cutoffs proposed by Schermelleh-Engel et al. (2003): comparative fit index (CFI), root mean square error approximation (RMSEA), standardized root mean square residual (SRMR), and χ^2 p value. A RMSEA value equal to or smaller than 0.05 indicates a good approximate model fit. The p value of the corresponding test of approximate fit should be equal to or smaller than 0.05. The CFI should be greater or equal to 0.97. Also, an SRMR greater than 0.05 and smaller than 0.1 would indicate an acceptable fit model, whereas values below 0.05 would indicate a good model fit.

We used two different estimators; robust maximum likelihood (MLR) was used for memory, and diagonally weighted least squares (WLSMV) for language and praxis items because the latter are ordinal observed variables (C.-H. Li, 2016). WLSMV estimator was also used for the three correlated-factor solution models where there are continuous and categorical indicators. Depending on the estimator, missing data were accommodated differently. Full-information maximum likelihood (invoked via MLR) yields consistent parameter estimates and standard errors when the missing data are missing-at-random (Rubin & Little, 2002). The WLSMV estimator uses pairwise variables present, meaning that each correlation is estimated using all available data. All analyses were conducted in M*plus* version 8.3 (Muthén & Muthén, 1998-2017).

Results

Our sample is constituted by 341 AD patients, 55% male, 75 years old on average (*SD* = 8 years old), and having 15 years of education (*SD* = 3). Major details, including population means on the MMSE and CDR (showing declines in cognitive and clinical status) over 24 months are shown in Table 1. Supplementary Tables 1 and 2 (available online) describe, for categorical items (i.e., Language and Praxis domains), proportions and counts at item-level over time. Supplementary Table 3 (available online) describes continuous items in terms of their minimum, maximum, mean, and standard deviation at item-level and across time. The Pearson correlation matrix between the items at each time point is shown in Figure 3. At the baseline evaluation, the pairwise correlations between items were weak, even within the same subdomains. The magnitude of the correlations did not appear to be uniform over time, increasing on each successive visit.

Table 2 shows the model fit indices for the configural model on the six unidimensional domains variations and in the two three-correlated factors solution specification for pre–post evaluation after 6 months. Some models showed nonpositive definite psi (Ψ) matrices—marked with "*"; this occurred due correlations higher than 1 between the latent variables, which implies an unacceptable solution. The same occurred for all models across four visits (Table 1).

Table 3 shows the model fit indices for the four visits after the inclusion of item-specific factors under different levels of invariance. In contrast to the classical invariance models (Table 2), language and memory subscales did not have nonpositive definite Ψ matrices after the inclusion of item-specific factors (Table 3). For these subscales and their variations the addition of item-specific factors also returned good fit models under at least configural specification; however, none of the models achieved scalar invariance, where thresholds (language subscale) and intercepts (memory) were held constant across time showed poor fit indices (Table 3). Praxis showed offending estimates even under a less restrictive solution.
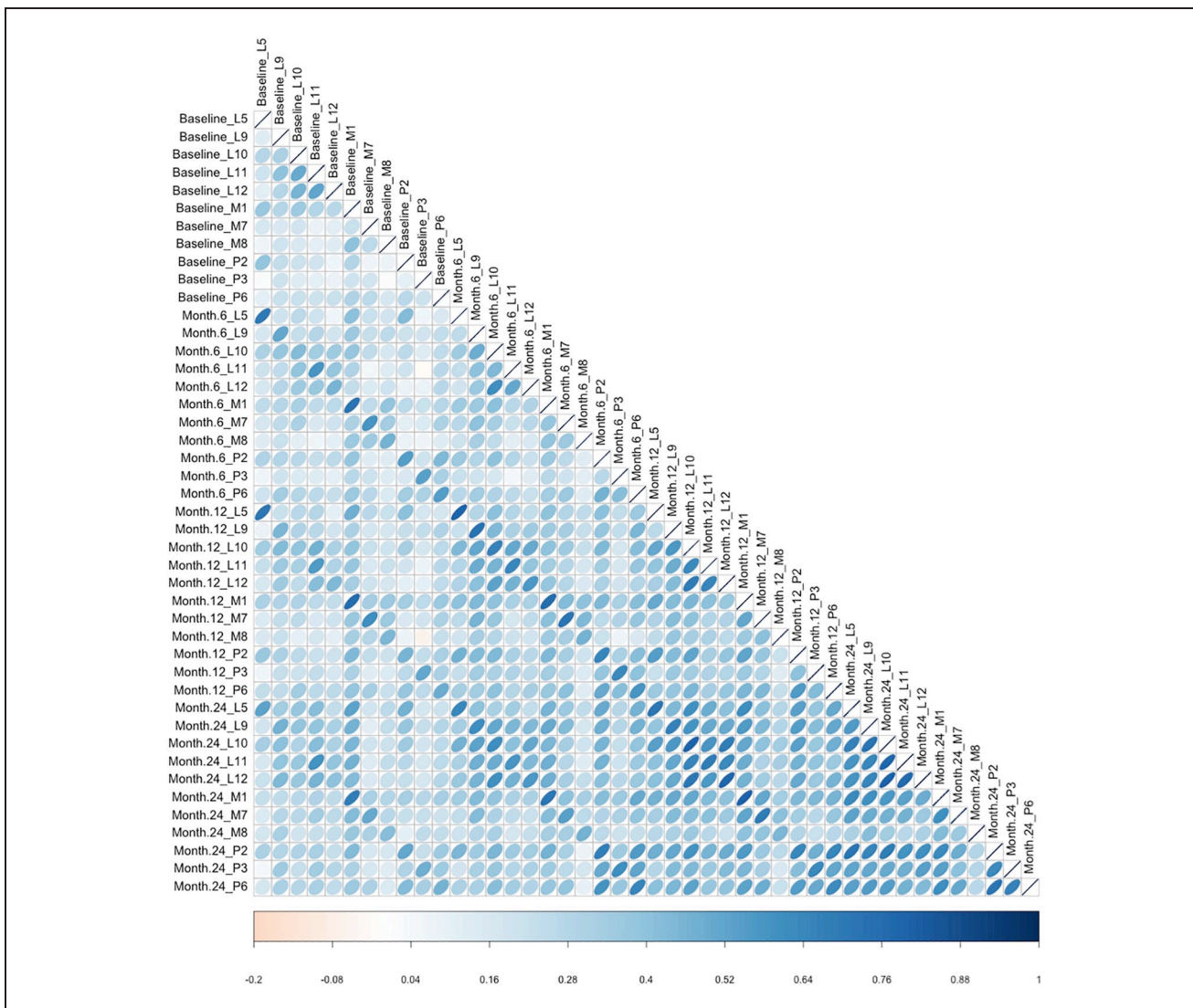
Figure 3. Pearson correlation matrix between all items at each time point.
Note. L = language items; M = memory items; P = praxis items.

The ADAS-Cog 11 and ADAS-Cog 13 memory subscale versions did not meet configural or weak invariance criteria. The memory subscale (Items 1, 4, 7, and 8, proposed to have construct validity in previous longitudinal analyses by Verma et al.), and the two language subscale configurations, returned admissible solutions and models with good fit indices under configural and weak restriction, although they did not meet scalar invariance.

The restrictions imposed from configural to weak invariance worsened the model for language with Items 2, 5, 10, 11, and 12 ($\chi^2_{(12)}$ difference test = 21.256, $p = .468$), but not for language with Items 5, 9, 10, 11, and 12 ($\chi^2_{(12)}$ difference test = 20.317, $p = .061$). For memory (Items 1, 4, 7, and 8) the restriction worsened the model (Sattora–Bentler Scaled chi-square difference $\chi^2_{(9)} = 20.200$, $p = .008$). For these

three best models, even though they did not achieve scalar invariance, we calculated the item reliability and item specificity across four visits. Table 4 shows the unstandardized and standardized factor loadings for the memory subscale (Items 1, 4, 7, and 8) and language subscales across the visits, and their reliability and item specificity.

For memory, the *word recognition* item was the least reliable over time (reliability ranged from 0.424 to 0.531), whereas the most reliable was *word recall* (average reliability 0.799). For the language subscale (Items 5, 9, 10, 11, and 12), the tasks had reliabilities higher than 0.69; *naming* was the most reliable (0.818) followed by *spoken language ability* (0.803), *word finding difficulty* (0.785), *remembering test instructions* (0.782), and *comprehension of spoken language* (0.725). For language (Items 2, 5, 10, 11, and 12), the

Table 2. Model Fit Indices for Traditional Configural Invariance Testing for Two and Four Waves of Assessment for Different ADAS-Cog and Subscale Specifications.

N_{waves}	N_{facts}	N_{items}	Nature of items	Scale or subdomain (Items or model)	χ^2	df	p	RMSEA	90% CI of RMSEA	Close fit	CFI	TLI	SRMR
2	2	10	Categorical	Language (5, 9, 10, 11, and 12)	364.647	34	<.001	0.169	[0.153, 0.185]	<.001	0.824	0.768	0.092
2	2	10	Categorical	Language (2, 5, 10, 11, and 12)	381.299	34	<.001	0.173	[0.158, 0.189]	<.001	0.807	0.745	0.107
2	2	8	Mixture	Memory (1, 7, 8, and 9)	138.982	19	<.001	0.136	[0.115, 0.158]	<.001	0.780	0.676	0.079
2	2	10	Mixture	Memory (1, 4, 7, 8, and 9)*	202.951	34	<.001	0.121	[0.105, 0.137]	<.001	0.763	0.686	0.082
2	2	8	Scalar	Memory (1, 4, 7, and 8)*	215.147	19	<.001	0.174	[0.153, 0.195]	<.001	0.773	0.666	0.074
2	2	6	Categorical	Praxis (2, 3, and 6)	121.721	8	<.001	0.204	[0.173, 0.237]	<.001	0.840	0.669	0.083
2	6	22	Mixture	Correlated-factor (ADAS-Cog 11)*	738.514	194	<.001	0.091	[0.084, 0.098]	<.001	0.793	0.753	0.090
2	6	24	Mixture	Correlated-factor (Verma et al., 2015)*	795.612	237	<.001	0.083	[0.077, 0.090]	<.001	0.815	0.784	0.087
4	4	20	Categorical	Language (5, 9, 10, 11, and 12)*	944.489	164	<.001	0.118	[0.111, 0.126]	<.001	0.882	0.864	0.114
4	4	20	Categorical	Language (2, 5, 10, 11, and 12)*	966.348	164	<.001	0.120	[0.113, 0.127]	<.001	0.882	0.863	0.119
4	4	16	Mixture	Memory (1, 7, 8, and 9)*	431.096	98	<.001	0.100	[0.090, 0.110]	<.001	0.738	0.679	0.097
4	4	20	Mixture	Memory (1, 4, 7, 8, and 9)*	674.507	164	<.001	0.096	[0.088, 0.103]	<.001	0.682	0.632	0.104
4	4	16	Scalar	Memory (1, 4, 7, and 8)*	811.504	98	<.001	0.146	[0.137, 0.156]	<.001	0.677	0.604	0.097
4	4	12	Categorical	Praxis*	367.415	48	<.001	0.140	[0.127, 0.153]	<.001	0.875	0.828	0.100
4	12	44	Mixture	Correlated-factor (ADAS-Cog 11)*	2,032.866	836	<.001	0.065	[0.061, 0.068]	<.001	0.837	0.0816	0.098
4	12	48	Mixture	Correlated-factor (Verma et al., 2015)*	2,253.780	1014	<.001	0.060	[0.057, 0.063]	<.001	0.849	0.832	0.099

Note. * = latent variable covariance matrix (psi) is not positive definite for this model due to correlation superior to 1; N_{waves} = number of waves; N_{facts} = number of factors being modelled across time; N_{items} = number of items being modelled across time; df = degrees of freedom; RMSEA = root mean square of error approximation; CI = confidence interval; CFI = comparative fit index; TLI =Tucker–Lewis index; SRMR = standardized root mean square residual. Item 1: word recall, Item 2: commands, Item 3: constructional praxis, Item 4: delayed word recall, Item 5: naming, Item 6: ideational praxis, Item 7: orientation, Item 8: word recognition, Item 9: remembering test instructions, Item 10: comprehension of spoken language, Item: 11 word finding difficulty, Item 12: spoken language ability, and Item 13: number cancellation. First row showed that the model for *language* invariance (formed by Items 5, 9, 10, 11, and 12) has two latent variables (one for baseline and other for 6 months); hence, there is a total of 10 manifest categorical items (i.e., five items of baseline loaded onto baseline factors and five items corresponding to after 6 months evaluation loaded on latent factor *6 months*).

Table 3. Invariance Testing for Models With Indicator-Specific Factors Under Four Assessments Across Time.

Model	Domain	N_{facts}	N_{items}	χ^2	df	p	RMSEA	90% CI of RMSEA	Close fit	CFI	TLI	SRMR
Configural	Language (5, 9, 10, 11, and 12)	8	20	178.101	142	.0216	0.027	[0.011, 0.039]	—	0.995	0.993	0.045
Weak	Language (5, 9, 10, 11, and 12)	8	20	197.984	154	.0097	0.029	[0.015, 0.040]	—	0.993	0.992	0.052
Strong	Language (5, 9, 10, 11, and 12)	8	20	644.570	221	<.001	0.075	[0.068, 0.082]	<0.001	0.936	0.945	0.057
Configural	Language (2, 5, 10, 11, and 12)	8	20	202.761	142	.0006	0.035	[0.024, 0.046]	0.990	0.991	0.988	0.048
Weak	Language (2, 5, 10, 11, and 12)	8	20	218.631	154	.0005	0.035	[0.024, 0.045]	0.993	0.991	0.988	0.054
Strong	Language (2, 5, 10, 11, and 12)	8	20	618.276	219	<.001	0.073	[0.066, 0.080]	<0.001	0.941	0.949	0.058
Configural	Memory (1, 7, 8, and 9)	7	16	125.502	83	.0018	0.039	[0.024, 0.052]	0.915	0.967	0.952	0.045
Weak	Memory (1, 7, 8, and 9)*	7	16	144.959	92	.0004	0.041	[0.028, 0.053]	0.876	0.958	0.946	0.054
Strong	Memory (1, 7, 8, and 9)*	7	16	484.908	116	<.001	0.097	[0.088, 0.106]	<0.001	0.710	0.700	0.086
Configural	Memory (1, 4, 7, 8, and 9)	8	20	175.841	142	.0283	0.026	[0.009, 0.038]	—	0.979	0.972	0.045
Weak	Memory (1, 4, 7, 8, and 9)*	8	20	209.977	154	.0018	0.033	[0.020, 0.043]	0.998	0.965	0.957	0.054
Strong	Memory (1, 4, 7, 8, and 9)*	8	20	548.413	181	<.001	0.077	[0.070, 0.085]	<0.001	0.771	0.760	0.082
Configural	Memory (1, 4, 7, 8)	7	16	133.957	83	.0003	0.042	[0.029, 0.055]	0.824	0.977	0.967	0.042
Weak	Memory (1, 4, 7, 8)	7	16	156.587	92	<.001	0.045	[0.033, 0.057]	0.725	0.971	0.962	0.061
Strong	Memory (1, 4, 7, 8)	7	16	449.206	104	<.001	0.099	[0.089, 0.108]	<0.001	0.844	0.820	0.180
Configural	Praxis*	6	12	37.852	39	.5521	0.000	[0.000, 0.036]	0.997	1.000	1.001	0.029
Weak	Praxis*	6	12	42.724	45	.5688	0.000	[0.000, 0.033]	0.999	1.000	1.001	0.032
Strong	Praxis*	6	12	269.673	83	<.001	0.081	[0.071, 0.092]	<0.001	0.927	0.942	0.041

Note. * = latent variable covariance matrix (psi) is not positive definite for this model due to correlation superior to 1. N_{facts} = number of factors being modelled; N_{items} = number of items being modelled; RMSEA = root mean square of error approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual.

Table 4. Standardized and Unstandardized Factor Solutions and Derived Reliabilities and Item Specificities for the Language and Memory Subscales in Models Disentangling Occasion and Indicator-Specific Information.

| Domain | Item | Visit | Loadings | | | | Reliability | Indicator specificity | Indicator specificity/ reliability |
| | | | Latent State Variable | | Indicator Specific Variable | | | | |
			Unst.	Stand.	Unst.	Stand.			
Memory[a]	1	Baseline	1.000	0.572	1.000	0.671	0.776	0.450	0.580
Memory[a]	4	Baseline	1.569	0.861			0.741		
Memory[a]	7	Baseline	0.696	0.240	1.000	0.592	0.466	0.350	0.752
Memory[a]	8	Baseline	1.309	0.385	1.000	0.530	0.429	0.281	0.655
Memory[a]	1	6 Months	1.000	0.556	1.014	0.685	0.779	0.469	0.602
Memory[a]	4	6 Months	1.661	0.811			0.658		
Memory[a]	7	6 Months	0.987	0.439	1.180	0.660	0.629	0.436	0.693
Memory[a]	8	6 Months	1.636	0.454	1.067	0.554	0.513	0.307	0.598
Memory[a]	1	12 Months	1.000	0.554	1.243	0.734	0.845	0.539	0.638
Memory[a]	4	12 Months	1.397	0.873			0.762		
Memory[a]	7	12 Months	1.085	0.501	1.481	0.754	0.820	0.569	0.693
Memory[a]	8	12 Months	1.260	0.392	1.018	0.520	0.424	0.270	0.638
Memory[a]	1	24 Months	1.000	0.539	1.201	0.713	0.799	0.508	0.636
Memory[a]	4	24 Months	1.269	0.916			0.839		
Memory[a]	7	24 Months	1.135	0.496	1.128	0.561	0.560	0.315	0.562
Memory[a]	8	24 Months	1.265	0.375	1.074	0.540	0.431	0.292	0.677
Language[a]	5	Baseline	1.000	0.503	1.000	0.709	0.755	0.503	0.666
Language[a]	9	Baseline	1.339	0.674	1.000	0.450	0.657	0.203	0.308
Language[a]	10	Baseline	1.481	0.745			0.556		
Language[a]	11	Baseline	1.454	0.732	1.000	0.495	0.781	0.245	0.314
Language[a]	12	Baseline	1.365	0.687	1.000	0.467	0.690	0.218	0.316
Language[a]	5	6 Months	1.000	0.493	1.071	0.759	0.819	0.576	0.703
Language[a]	9	6 Months	1.458	0.719	1.348	0.607	0.886	0.368	0.416
Language[a]	10	6 Months	1.743	0.860			0.739		
Language[a]	11	6 Months	1.308	0.645	1.176	0.582	0.755	0.339	0.449
Language[a]	12	6 Months	1.588	0.783	0.809	0.377	0.756	0.142	0.188
Language[a]	5	12 Months	1.000	0.593	1.060	0.751	0.915	0.564	0.616
Language[a]	9	12 Months	1.173	0.695	1.346	0.606	0.851	0.367	0.432
Language[a]	10	12 Months	1.603	0.950			0.903		
Language[a]	11	12 Months	1.238	0.734	0.911	0.451	0.742	0.203	0.274
Language[a]	12	12 Months	1.435	0.851	0.884	0.412	0.894	0.170	0.190
Language[a]	5	24 Months	1.000	0.718	0.734	0.520	0.786	0.270	0.344
Language[a]	9	24 Months	1.044	0.750	0.919	0.414	0.734	0.171	0.234
Language[a]	10	24 Months	1.295	0.930			0.865		
Language[a]	11	24 Months	1.160	0.833	0.832	0.412	0.863	0.170	0.197
Language[a]	12	24 Months	1.247	0.896	0.578	0.270	0.875	0.073	0.083
Language[b]	2	Baseline	1.000	0.376	1.000	0.669	0.588	0.448	0.761
Language[b]	5	Baseline	1.369	0.515	1.000	0.702	0.758	0.493	0.650
Language[b]	10	Baseline	2.013	0.756			0.572		
Language[b]	11	Baseline	1.930	0.725	1.000	0.531	0.808	0.282	0.349
Language[b]	12	Baseline	1.878	0.706	1.000	0.440	0.692	0.194	0.280
Language[b]	2	6 Months	1	0.531	0.981	0.656	0.712	0.430	0.604
Language[b]	5	6 Months	0.888	0.471	1.080	0.759	0.798	0.576	0.722
Language[b]	10	6 Months	1.602	0.850			0.722		
Language[b]	11	6 Months	1.221	0.648	1.066	0.566	0.740	0.320	0.433
Language[b]	12	6 Months	1.452	0.770	0.921	0.406	0.758	0.165	0.217
Language[b]	2	12 Months	1.000	0.551	0.875	0.585	0.647	0.342	0.529

(continued)

Table 4. (continued)

Domain	Item	Visit	Loadings				Reliability	Indicator specificity	Indicator specificity/ reliability
			Latent State Variable		Indicator Specific Variable				
			Unst.	Stand.	Unst.	Stand.			
Language[b]	5	12 Months	1.070	0.590	1.077	0.756	0.920	0.572	0.621
Language[b]	10	12 Months	1.724	0.951			0.904		
Language[b]	11	12 Months	1.319	0.727	0.869	0.461	0.742	0.213	0.286
Language[b]	12	12 Months	1.557	0.859	0.910	0.401	0.898	0.161	0.179
Language[b]	2	24 Months	1.000	0.683	0.761	0.509	0.726	0,.259	0.357
Language[b]	5	24 Months	1.041	0.711	0.780	0.548	0.806	0.300	0.373
Language[b]	10	24 Months	1.374	0.939			0.882		
Language[b]	11	24 Months	1.225	0.837	0.747	0.397	0.857	0.158	0.184
Language[b]	12	24 Months	1.307	0.893	0.623	0.275	0.873	0.076	0.087

Note. Uns. = unstandardized; stand. = standardized.
[a]Specified as per (Verma, Beretvas, Pascual, Masdeu and Markey, 2015).
[b]Specified as per the ADAS-Cog 11 and ADAS-Cog 13.

highest reliability on average was also that of *naming* and the lowest was that of *commands*.

The majority of variance in the memory subscale (Items 1, 4, 7, and 8) was due to item specificity (ranging from 56.2% to 75.2%). Because *delayed word recall* was chosen as reference, the latent state variables are the true state variable of delayed word recall. The indicator-specific factors of the other indicators represent that part of the indicator that cannot be predicted by the reference indicator and is, therefore, not shared with the reference indicator (Eid & Kutscher, 2014). For example, 75.2% of the reliable information provided by the *orientation* task at the first assessment (i.e., 0.35/0.466) corresponded to the true state score of *orientation* that was not shared with the other indicators. For language, using *comprehension of spoken language* as the reference indicator, we observed more heterogeneity in the specificities of the items (ranging from 0.073 for *spoken language ability* to 0.576 for *naming*).

Supplementary Table 4 (available online) shows the standardized Pearson correlations between the indicator-specific latent variables and the correlations of the latent state variables, also known as latent stability coefficients (see Eid & Kutscher 2014). For memory, the correlations between the indicator-specific variables were in the majority low–moderate, ranging from 0.318 to 0.445. For language, they ranged from 0.130 to 0.538. The correlations of the latent state variables were in the majority strong, ranging from 0.771 to 0.878 for memory and from 0.671 to 0.870. On this measure, higher correlations indicate smaller interindividual differences in intraindividual change. In terms of attrition, the lowest covariance coverage (i.e., percentage of people with nonmissing values on the variable or pair of variables) was 47.2%, which is more than minimum covariance coverage of 10% used by M*plus* to interrupt the analysis (see Supplementary Table 5, available online).

Discussion

Longitudinal invariance testing is commonly conducted in psychological assessments (Chan et al., 2015; McFall et al., 2015), but it has been applied scarcely to neurological instruments, limiting our comprehension about whether those tools track progression accurately when researchers use their summed scores in ordinary statistical testing. Among mild AD patients, under traditional invariance testing, even configural invariance, which is the most fundamental and least restrictive level of invariance (Cheung & Rensvold, 2002; Eid & Kutscher, 2014), was not achieved for the ADAS-Cog 11 structure, or for an updated factor structure based on item response theory, under a three correlated factors solution. The majority of the models (identified specifically in Table 2) showed correlations between the latent factors higher than one, which are not admissible; even the simplest models with only two latent variables were problematic. After adding item-specific effects, two language subdomain models and one memory subdomain model achieved configural and metric invariance across four time points. In these cases, rank-order comparisons (e.g., between subjects or within subjects over time) might be valid; however, little support for scalar invariance was achieved, suggesting that mean-level comparisons might be discouraged even for these subscales.

A necessary condition for a CFA solution is that both the input variance—covariance matrix (i.e., coming from the data) and the model-implied variance—covariance matrix (i.e., based on the two ways the models were specified) are positive definite. The improper empirical solutions obtained here suggest that the specified models are very different from the structure that the data would support.

Failure to achieve longitudinal configural invariance casts doubt on the adequacy of the ADAS-Cog general

score and subdomain scores to capture meaningful changes over time. Without meeting invariance criteria, the assumptions underlying conventional statistical approaches (e.g., paired *t* tests, repeated measures analyses of variance, generalized estimating equations, etc.) are violated, and comparing means over time may be misleading (Cheung & Rensvold, 2002). The ADAS-Cog score and estimates of its trajectory might be biased because the covariances between the items are not equal from visit to visit. In other words, the association between the items (test scores) and the latent factor (performance on that occasion) are inconsistent over time. Moreover, the reliable variance (the part of the score that is truly related to a person's performance on a given visit) is also not consistent between visits. These problems may lead to differences in ADAS-Cog scores from one visit to the next that do not reflect the magnitude of the change in the underlying feature(s) that the ADAS-Cog is intended to measure.

Only one study to our knowledge mentioned the use of longitudinal invariance of the ADAS-Cog using data from ADNI (Dowling et al., 2016). The results from longitudinal invariance were not shown explicitly; however, it might be noted that a maximum likelihood estimator appears to have been used for the main analysis of the article (Akaike information criterion, Bayesian Information criterion were presented), so it might intuited that the ordered-categorical items of ADAS-Cog (e.g., those from language and praxis subscales) were treated as continuous variables, which can lead to bias in the estimation of the factor loadings (Rhemtulla et al., 2012). Moreover, some of the analyses of Dowling et al. (2016) were conducted using a complete-cases approach, which requires strong assumptions regarding a missing completely at random mechanism (White & Carlin, 2010), and which might generate selection bias. In that study, the model of change fit indices might be considered to indicate that further adaptations in model specification would be beneficial (Schermelleh-Engel et al., 2003). In agreement, we hypothesized that item-specific effects may play an important role.

To deal with the configural issues in the traditional models, the addition of item-specific factors into the language and memory models returned models that fit the data well; however, they failed to achieve strong invariance. The present findings empirically demonstrate that each item, even within the same subdomain, works differently, having an important amount item-specific variance. In other words, neither the memory nor language items reflected uniform constructs, but rather they captured distinct components that progress differently in AD. The small correlations observed between the items (seen visually in Figure 3) indicate that they share little information, consistent with the strong item-specific effects observed. Over time, the correlations between the items increased, albeit to at most a moderate correlation at the final visit; the increase in the

correlation across time is often attributable to *practice effects* (Duff, 2012), which have been described in people with mild cognitive impairment (Calamia et al., 2012). The results *in toto* suggest that even within the same subdomain, the items did not measure the same phenomenon. In this context, the principle of aggregation (Rushton et al., 1983), where the sum of the items would have greater reliability than any individual item alone, is not applicable. Instead, summing to create an aggregate score will result in a considerable loss of information, and it may introduce unintended bias. Statistically, the different scaling features (i.e., ordinal items for language and praxis vs. continuous measures for memory items) pose conceptual problems in creating a general score from their sums, which also argues that they might be best considered individually.

Our second set of models (i.e., those taking into account item-specific variances), supported reasonable reliabilities for many of the ADAS-Cog items; however, those models revealed that the majority of the reliable variance in memory was accounted for by the use of the same items over time, rather than the construct that the items intended to measure at each visit, casting doubt on their ability to track meaningful change over time. For instance memory Items 1 and 4 had reliabilities consistently greater than 0.6 across the four visits, but less than half of that variance was related to change in the overall construct of memory over time. The reliabilities of the language items were generally acceptable (i.e., higher than 0.7) with the exception of Items 9, 10, and 12 at the baseline assessment; however, the amount of information specific to each item was inconsistent over time (e.g., for Item 12, from baseline to the fourth evaluation, it declined from 32% to 8% and for Item 5, which had more item-specific information at baseline, it declined from 67% to 34% at the fourth visit), and Items 5, 9, and 11 increased their item-specific variance from baseline to the second visit. In examining indicator-specific effects, very low latent variable correlations in the matrix of indicator-specific variables indicated that there were small differences between people in their trajectories over time using the language and memory subscales. For this and the reasons aforementioned, general aggregate ADAS-Cog or subscale scores (summing the items) might be discouraged when tracking change over time. Instead, the most relevant individual items might be evaluated separately over repeated assessments.

Considering each of the items individually would preserve the information they contain, which may be more reliable than total scores derived from their sum; but even so, for any given item, the amount of information that is related to a person's enduring/consistent capabilities versus their trajectory of decline in performance or effort at a given visit might be uncertain. In a practical sense, a prespecified single item endpoint might be chosen or correction for multiple comparisons (e.g., multivariate analysis of covariance or

false discovery rate correction) might be necessary to control Type I error. Here, we modelled effects specific to each item over time, but still more sophisticated models might estimate trait change, occasion-specific effects, and accumulated situational effects simultaneously in order to more specifically disentangle the information that is related to a person's cognitive trajectory (Eid et al., 2017).

In the field of AD, there is considerable interest in identifying sensitive and meaningful outcomes for monitoring, and for use in clinical trials, including surrogate biofluid biomarkers, neuroimaging volumetrics, neuropsychological measures, activities of daily living scales, neurophysiological/functional brain parameters, amyloid in the lens of the eye, ecological functional outcomes, or actigraphy data (Frisoni et al., 2019), an approach consistent with the amyloid-tau-neurodegeneration framework proposed by Jack et al. (2018). It has been suggested that aggregation of multiple measures within or across these domains might be useful to eliminate the noise inherent in each individual measure, resulting in more robust or precise estimates of the trajectory of decline. Including different measurements (biomarker, imaging, and scales) adds a new type of information that might be controlled for by regression of method effects. Different models are available to disentangle trait features and methodological features, (Eid et al., 2003; Eid et al., 2008) which might be extended into longitudinal designs for this purpose (Geiser et al., 2010). We and others have found reliable aggregate estimates via CFA for different types of measures in AD (e.g., inflammatory/immunological; Swardfager et al., 2017; Bawa et al., 2020), but the behavior of these measures over time has not been evaluated. The principles and methods applied here might also be considered in the development of composite surrogate trial outcome measures, to gain some insight into how repeated measures using those instruments would track change prior to their use in trials.

As a potential limitation, the data examined here describe the natural course of AD over 6 or 24 months; however, some clinical trials involve more repeated measures over shorter outcome windows. To establish generalizability in that context, evaluation of repeated measures trial data over a shorter timeline would be useful; however, in the present study, the instrument failed to achieve basic invariance between two measurements over 6 months among 341 patients with mild AD consistent with the size and duration of many clinical trials for AD (Birks & Harvey, 2018; McShane et al., 2019), suggesting that using this instrument even over short windows of observation could be problematic. Because the second, more flexible modeling approach assumes a unidimensional structure, we were unable to apply it to the ADAS-Cog three-correlated factors models because such complexity increases offending estimates such as those observed for the total scale, even over two visits 6 months apart. As a further limitation, we did not conduct partial scalar invariance testing (Van de Schoot et al., 2012). Although that approach might improve model fit indices, perhaps achieving scalar invariance for the Eid et al.'s (2014; Eid et al., 2016) models, improvement in the fit of these models would not dismiss the marked heterogeneous item-specific effects observed. Last, because the different natures of items necessitated different estimators (i.e., for categorical items, WLSMV; for continuous, MLR; and for mixtures of continuous and ordered-categorical items, WLSMV), missing data have been handled differently, and different mechanisms of missingness have been assumed.

Conclusion

Using a large dataset from an observational study, we failed to provide evidence that the ADAS-Cog or its subscales achieved invariance in mild AD over time frames of 6 or 24 months. As dementia in this population progressed over time, the measurement properties of the ADAS-Cog also changed. Achieving scalar invariance is the minimal requirement to allow ordinary mean comparisons over time using procedures such as paired *t* tests or repeated measures ANOVA. Applying models that relaxed the assumption that the trajectories of the items would be stable or uniform revealed heterogeneous item-specific effects, some of which were strong. For this reason, mean comparisons of their sums in statistical tests is likely to produce unreliable inferences. The use of ADAS-Cog total or subscale scores to track trajectories of cognitive decline over time might be reconsidered.

Author's Note

Saffire H. Krance is also affiliated with the Western University, London, Ontario, Canada.

Declaration of Conflicting Interests

Funding

ORCID iDs

Hugo Cogo-Moreira [iD] https://orcid.org/0000-0001-9411-9237
Saffire H. Krance [iD] https://orcid.org/0000-0001-6679-4067

Supplemental Material

Supplemental material for this article is available online.

References

Arevalo-Rodriguez, I., Smailagic, N., Roqué i Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., Pedraza, O. L., Cosp, X. B., & Cullum, S. (2015). Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, *2015*(3), Article CD01078. https://doi.org/10.1002/14651858.CD010783.pub2

Bawa, K. K., Krance, S. H., Herrmann, N., Cogo-Moreira, H., Ouk, M., Yu, D., Wu. C. Y., Black, S. E., Lanctôt. K. L., Swardfager, W.; for the Alzheimer's Disease Neuroimaging Initiative. (2020). A peripheral neutrophil-related inflammatory factor predicts a decline in executive function in mild Alzheimer's disease. *Journal of Neuroinflammation*, *17*(84), 1-11. https://doi.org/10.1186/s12974-020-01750-3

Birks, J. S., & Harvey, R. J. (2018). Donepezil for dementia due to Alzheimer's disease. *Cochrane Database Systematic Review*, *2018*(6), Article CD001190. https://doi.org/10.1002/14651858.CD001190.pub3

Bollen, K. A. (1998). *Structural equations with latent variables*. Wiley.

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

Cano, S. J., Posner, H. B., Moline, M. L., Hurt, S. W., Swartz, J., Hsu, T., & Hobart, J. C. (2010). The ADAS-cog in Alzheimer's disease clinical trials: Psychometric evaluation of the sum and its parts. *Journal of Neurology, Neurosurgical & Psychiatry*, *81*(12), 1363-1368. https://doi.org/10.1136/jnnp.2009.204008

Chan, R. C. K., Dai, S., Lui, S. S. Y., Ho, K. K. Y., Hung, K. S. Y., Wang, Y., Geng, F., Li, Z., & Cheung, E. F. C. (2015). Re-visiting the nature and relationships between neurological signs and neurocognitive functions in first-episode schizophrenia: An invariance model across time. *Scientific Reports*, *5*, Article 11850. https://doi.org/10.1038/srep11850

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

Connor, D. J., & Sabbagh, M. N. (2008). Administration and scoring variance on the ADAS-Cog. *Journal of Alzheimer's Disease*, *15*(3), 461-464. https://doi.org/10.3233/JAD-2008-15312

Dowling, N. M., Bolt, D. M., & Deng, S. (2016). An approach for estimating item sensitivity to within-person change over time: An illustration using the Alzheimer's Disease Assessment Scale–Cognitive subscale (ADAS-Cog). *Psychological Assessment*, *28*(12), 1576-1585. https://doi.org/10.1037/pas0000285

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(3), 248-261. https://doi.org/10.1093/arclin/acr120

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research*, *1*(4), 55-65. https://www.dgps.de/fachgruppen/methoden/mpr-online/issue1/art4/eid.pdf

Eid, M., Geiser, C., & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science*, *25*(4), 275-280. https://doi.org/10.1177/0963721416649624

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects. *European Journal of Psychological Assessment*, *33*(4), 285-295. https://doi.org/10.1027/1015-5759/a000435

Eid, M., & Kutscher, T. (2014). Statistical models for analyzing stability and change in happiness. In K. M. Sheldon, & R. E. Lucas (Eds.), *Stability of happiness* (pp. 261-297). Academic Press.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator

CT-C (M-1) model. *Psychological Methods*, *8*(1), 38-60. https://doi.org/10.1037/1082-989X.8.1.38

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*(3), 230-253. https://doi.org/10.1037/a0013219

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, *61*(1), 50-55. https://doi.org/10.1037/0003-066X.61.1.50

Frisoni, G. B., Blin, O., & Bordet, R. (2019). One step forward toward a surrogate endpoint for clinical trials of Alzheimer's disease drugs: The results of PharmaCog WP5 (European ADNI). *Journal of Alzheimer's Disease*, *69*(1), 1-2. https://doi.org/10.3233/jad-190267

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multi-method change model to depression and anxiety in children. *Developmental Psychology*, *46*(1), 29-45. https://doi.org/10.1037/a0017888

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, *17*(2), 255-283. https://doi.org/10.1037/a0026977

Grochowalski, J. H., Liu, Y., & Siedlecki, K. L. (2016). Examining the reliability of ADAS-Cog change scores. *Aging Neuropsychology and Cognition*, *23*(5), 513-529. https://doi.org/10.1080/13825585.2015.1127320

Hobart, J., Cano, S., Posner, H., Selnes, O., Stern, Y., Thomas, R., & Zajicek, J. (2013). Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods. *Alzheimer's & Dementia*, *9*(15), S4-S9. https://doi.org/10.1016/j.jalz.2012.08.005

Honig, L. S., Vellas, B., Woodward, M., Boada, M., Bullock, R., Borrie, M., Hager, K., Andreasen, N., Scarpini, E., Liu-Seifert, H., Case, M., Dean, R. A., Hake, A., Sundell, K., Hoffmann, V. P., Carlson, C., Khanna, R., Mintun, M., DeMattos, R., . . .Siemers, E. (2018). Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. *New England Journal of Medicine*, *378*(4), 321-330. https://doi.org/10.1056/NEJMoa1705971

Horn, J. L., & McArdle, J. J. (1992). A practical and theo-retical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117-144. https://doi.org/10.1080/03610739208253916

Jack, C. R., Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., & Sperling, S. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535-562. https://doi.org/10.1016/j.jalz.2018.02.018

Karin, A., Hannesdottir, K., Jaeger, J., Annas, P., Segerdahl, M., Karlsson, P., Sjögren, N., von Rosen, T., & Miller, F. (2014). Psychometric evaluation of ADAS-Cog and NTB for measur-ing drug response. *Acta Neurologica Scandinavica*, *129*(2), 114-122. https://doi.org/10.1111/ane.12153

Kueper, J. K., Speechley, M., & Montero-Odasso, M. (2018). The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and responsiveness in pre-dementia populations: A narrative review. *Journal of Alzheimer's Disease*, *63*(2), 423-444. https://doi.org/10.3233/jad-170991

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Li, D.-D., Zhang, Y.-H., Zhang, W., & Zhao, P. (2019). Meta-analysis of randomized controlled trials on the efficacy and safety of Donepezil, Galantamine, Rivastigmine, and Memantine for the treatment of Alzheimer's disease. *Frontiers in Neuroscience*, *13*, Article 472. https://doi.org/10.3389/fnins.2019.00472

McFall, G. P., Wiebe, S. A., Vergote, D., Westaway, D., Jhamandas, J., Backman, L., & Dixon, R. A. (2015). ApoE and pulse pressure interactively influence level and change in the aging of episodic memory: Protective effects among epsilon2 carriers. *Neuropsychology*, *29*(3), 388-401. https://doi.org/10.1037/neu0000150

McShane, R., Westby, M. J., Roberts, E., Minakaran, N., Schneider, L., Farrimond, L. E., Ware, J., & Debarros, J. (2019). Memantine for dementia. *Cochrane Database Systematic Review*, *3*, Article CD003154. https://doi.org/10.1002/14651858.CD003154.pub6

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/bf02294825

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). American Psychological Association.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*: Routledge.

Mohs, R. C., Marin, D., Green, C. R., & Davis, K. L. (1997). The Alzheimer's Disease Assessment Scale: Modifications that can enhance its use in future clinical trials. In R. E. Becker, E. Giacobini, J. M. Barton, & M. Brown (Eds.), *Alzheimer disease: Advances in Alzheimer disease therapy* (pp. 407-411). Birkhäuser.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A com-parison of robust continuous and categorical SEM estima-tion methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373. https://doi.org/10.1037/a0029315

Rockwood, K., Fay, S., Gorman, M., Carver, D., & Graham, J. E. (2007). The clinical meaningfulness of ADAS-Cog changes in Alzheimer's disease patients treated with donepezil in an open-label trial. *BMC Neurology*, *7*, Article 26. https://doi.org/10.1186/1471-2377-7-26

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

Rubin, D. B., & Little, R. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*(1), 18-38. https://doi.org/10.1037/0033-2909.94.1.18

Salloway, S., Sperling, R., Fox, N. C., Blennow, K., Klunk, W., Raskind, M., Sabbagh, M., Honig, L. S., Porsteinsson, A. P., Ferris, S., Reichert, M., Ketter, N., Nejadnik, B., Guenzler, V., Miloslavsky, M., Wang, D., Lu, Y., Lull, J., Tudor, J. C., . . .Brashear, H. R. (2014). Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *New England Journal of Medicine*, *370*(4), 322-333. https://doi.org/10.1056/NEJMoa1304839

Saunders, A. M., Strittmatter, W. J., Schmechel, D., George-Hyslop, P. S., Pericak-Vance, M. A., Joo, S., Rosi, B. L., Gusella, J. F., Crapper-MacLachlan, D. R., Alberts, M. J., Hulette, C., Crain, B., Goldgaber, D., & Roses, A. D. (1993). Association of apolipoprotein E allele $\in$4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, *43*(8), 1467-1467. https://doi.org/10.1212/WNL.43.8.1467

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23-74. https://www.researchgate.net/publication/251060246_Evaluating_the_Fit_of_Structural_Equation_Models_Tests_of_Significance_and_Descriptive_Goodness-of-Fit_Measures

Swardfager, W., Yu, D., Ramirez, J., Cogo-Moreira, H., Szilagyi, G., Holmes, M. F., Scott, C. J. M., Scola, G., Chan, P. C., Chen, J., Chan, P., Sahlas, D. J., Herrmann, N., Lanctôt, K. L., Andreazza, A. C., Pettersen, J. A., & Black, S. E. (2017). Peripheral inflammatory markers indicate microstructural damage within periventricular white matter hyperintensities in Alzheimer's disease: A preliminary report. *Alzheimer's & Dementia*, *7*(1), 56-60. https://doi.org/10.1016/j.dadm.2016.12.011

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492.

Van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, *6*, Article 1064. https://doi.org/10.3389/fpsyg.2015.01064

Verma, N., Beretvas, S. N., Pascual, B., Masdeu, J. C., Markey, M. K., & Alzheimer's Disease Neuroimaging Initiative. (2015). New scoring methodology improves the sensitivity of the Alzheimer's Disease Assessment Scale–Cognitive subscale (ADAS-Cog) in clinical trials. *Alzheimer's Research & Therapy*, *7*(1), Article 64. https://doi.org/10.1186/s13195-015-0151-0

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., Liu, E., Morris, J. C., Petersen, R. C., Saykin, A. J., Schmidt, M. E., Shaw, L., Siuciak, J. A., Soares, H., Toga, A. W., & Trojanowski, J. Q. (2013). The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, *9*(5), e111-e194. https://doi.org/10.1016/j.jalz.2013.05.1769

Weyer, G., Erzigkeit, H., Kanowski, S., Ihl, R., & Hadler, D. (1997). Alzheimer's Disease Assessment Scale: Reliability and validity in a multicenter clinical trial. *International Psychogeriatrics*, *9*(2), 123-138. https://doi.org/10.1017/S1041610297004298

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *29*(28), 2920-2931. https://doi.org/10.1002/sim.3944

Yagi, T., Kanekiyo, M., Ito, J., Ihara, R., Suzuki, K., Iwata, A., Iwatsubo, T., Aoshima, K., & Alzheimer's Disease Neuroimaging Initiative, & Japanese Alzheimer's Disease Neuroimaging Initiative. (2019). Identification of prognostic factors to predict cognitive decline of patients with early Alzheimer's disease in the Japanese Alzheimer's Disease Neuroimaging Initiative study. *Alzheimer's & Dementia*, *5*(3), 364-373. https://doi.org/10.1016/j.trci.2019.06.004