the british psychological society
promoting excellence in psychology

Journal of Neuropsychology (2022), 16, 324–352

www.wileyonlinelibrary.com

Research Article

Measurement invariance testing of longitudinal neuropsychiatric test scores distinguishes pathological from normative cognitive decline and highlights its potential in early detection research

Sophia Haberstumpf*†[1], André Forster†[2], Jonas Leinweber[3], Stefanie Rauskolb[4], Johannes Hewig[2], Michael Sendtner[4], Martin Lauer[1], Thomas Polak[1], Jürgen Deckert[1] and Martin J. Herrmann[1]

[1]Center for Mental Health, Department of Psychiatry, Psychosomatics and Psychotherapy, University Hospital Würzburg, Würzburg, Germany
[2]Institute of Psychology, Julius-Maximilians-University, Würzburg, Germany
[3]Alexianer Psychiatric Hospital Köln-Porz, Köln, Germany
[4]Institute of Clinical Neurobiology, University Hospital Würzburg, Würzburg, Germany

Objective. Alzheimer's disease (AD) is a growing challenge worldwide, which is why the search for early-onset predictors must be focused as soon as possible. Longitudinal studies that investigate courses of neuropsychological and other variables screen for such predictors correlated to mild cognitive impairment (MCI). However, one often neglected issue in analyses of such studies is measurement invariance (MI), which is often assumed but not tested for. This study uses the absence of MI (non-MI) and latent factor scores instead of composite variables to assess properties of cognitive domains, compensation mechanisms, and their predictability to establish a method for a more comprehensive understanding of pathological cognitive decline.

Methods. An exploratory factor analysis (EFA) and a set of increasingly restricted confirmatory factor analyses (CFAs) were conducted to find latent factors, compared them with the composite approach, and to test for longitudinal (partial-)MI in a neuropsychiatric test battery, consisting of 14 test variables. A total of 330 elderly (mean age: 73.78 ± 1.52 years at baseline) were analyzed two times (3 years apart).

Results. EFA revealed a four-factor model representing declarative memory, attention, working memory, and visual–spatial processing. Based on CFA, an accurate model was

**Correspondence should be addressed to Sophia Haberstumpf, University Hospital Würzburg, Center of Mental Health, Department of Psychiatry, Psychosomatics and Psychotherapy, Margarete-Höppel-Platz 1, D - 97080 Würzburg, Germany (email: haberstump_s@ukw.de).*
†*Both authors contributed equally to this work.*

DOI:10.1111/jnp.12269

estimated across both measurement timepoints. Partial non-MI was found for parameters such as loadings, test- and latent factor intercepts as well as latent factor variances. The latent factor approach was preferable to the composite approach.

Conclusion. The overall assessment of non-MI latent factors may pose a possible target for this field of research. Hence, the non-MI of variances indicated variables that are especially suited for the prediction of pathological cognitive decline, while non-MI of intercepts indicated general aging-related decline. As a result, the sole assessment of MI may help distinguish pathological from normative aging processes and additionally may reveal compensatory neuropsychological mechanisms.

Mild cognitive impairment (MCI) and Alzheimer's disease (AD)

Due to the constantly aging society, neurodegenerative diseases such as dementia represent a growing challenge for health care systems worldwide (Abbott, 2011; Bickel, 2001; Prince et al., 2013). Affecting 60–70% of people suffering from dementia, one of the most usual forms is Alzheimer's disease (AD; World Health Organization [WHO], 2016). An early indicator is mild cognitive impairment (MCI), which often progresses into AD (Arnáiz & Almkvist, 2003). According to Petersen (2000), 10–15% of MCI patients convert into AD per year, and up to 15–20% of the general population express MCI symptomatology. Even though there is no cure available to date, early interventions can dampen the course of the disease (Mayeux, 2010; Winblad et al., 2006), which highlights the necessity of diagnostics in the early stages. Thus, finding variables with high predictive power for neuropsychiatric changes is a focus of MCI-related research.

According to the consensus formulated in DSM-V and ICD-10 (American Psychiatric Association, 2014; World Health Organization [WHO], 2019), diagnostics of MCI and AD that heavily rely on neuropsychiatric tests as their first symptoms are deficits in cognitive performance, such as memory loss (Arnáiz & Almkvist, 2003; Jahn, 2013; Nestor, Fryer, & Hodges, 2006; Riedel & Blokland, 2015). As a result, finding predictors for such neuropsychological symptoms may elude targets for early interventions. The statistically and methodologically most efficient way to address this topic is analyzing longitudinal within-subject course data (Cooper, Sommerlad, Lyketsos, & Livingston, 2015; Hendrix et al., 2015; Makkar et al., 2020).

Shortcomings of the composite approach

A valid approach to increase robustness and significance of prediction analyses may be to create composite variables consisting of a sum or average score of potential predictors of interest. For example, multiple performance scores can be combined by forming a composite score. However, by simply adding the test scores, it is implicitly assumed that all scores are equally meaningful for the target construct (e.g., declarative memory). However, since the target construct is often a latent factor, it should be empirically verified that this assumption is, indeed, true. To do so, however, a latent factor approach would be more adequate. This problem is further complicated by the fact that the extent to which a predictor is relevant to the latent construct can vary across groups and over time. As a result, both the classical composite and weighted composite approaches that impose fixed weights on scores within the composite (e.g., $1 \times$ immediate memory performance $+ 0.3 \times$ working memory performance = latent memory ability) may fall short if the actual relationship of the manifest test scores differs from the weights chosen by the researcher (in the classical approach, each variable is multiplied by a weight of 1). Factor analyses may provide the most reliable weights for calculating composites. This

may be particularly the case in longitudinal studies, as weights may change over time, which may affect the comparability of measurement occasions within the follow-up data. This effect ("response shift") has been described in other areas of research (e.g., Oort, 2005). However, different weights are not the only parameters that can change over time, which further complicates analyses and suggests new ways to examine course data in detail. For instance, if a sample achieved a mean score of 10 on a composite variable described by researchers as an indicator of memory at both the first and second measurement occasions, one would conclude that the sample's memory performance had not changed. However, this null finding could be misleading, as this sample's latent declarative memory performance may have decreased, even if this did not manifest in the composite variable due to compensatory mechanisms (e.g., coping strategies, test-memory effects). Thus, to estimate latent ability changes and to detect effect-concealing or effect-inflating mechanisms, the intercorrelation matrix of different neuropsychiatric tests can be used. For example, an altered covariation between memory and attention scores at the second measurement occasion may indicate that the ability to modulate attention might cause a decline in memory performance less noticeable. Also, memory abilities might have a lower covariance with other latent skills if its scores were affected by retest effects, while other neuropsychiatric domains were not. Hence, merit lies in the analysis of test score interplays rather than absolute values. The classical method to deal with such complex matrices between multiple test scores is the (confirmatory) factor analysis, which extracts latent abilities from manifest test scores and estimates changes in the intercorrelation of manifest and latent variables based on these data. In summary, this approach investigates the equivalence of parameters within a structural equation model (SEM) across groups/time and can find indicators of possible bias mechanisms that may distort the results of the composite approach. Measurement invariance (MI; no significant variation of a parameter across groups/time) of parameters would imply that the manifest sum score approach would be largely unbiased. The following section gives interpretations for non-invariance for a subset of central parameters within such analyses.

Longitudinal MI

In most studies investigating MI, SEM comparing increasingly restricted confirmatory factor models is the method of choice. Due to its ability to integrate latent and observed variables out of many test variables, this approach is expected to offer an appropriate statistical method to reveal latent factor structures and to prove construct validity by factorial invariance measurements of neuropsychiatric test batteries across time, sample subgroups, and different cognitive levels (Berndt & Williams, 2013; Kline, 2005; Mungas, Widaman, Reed, & Tomaszewski Farias, 2011; Park & Festini, 2017; Rahmadi et al., 2018; Rowe, 2010; Schumacker & Lomax, 2004).

Intercepts

One often recognised MI parameter is the estimated intercept of single items/tests and latent means. In the context of regression (which reflects the relationship of a latent factor to its manifest indicators), intercepts reflect the (grand) mean score of a given population. Non-MI, for example, increase/decrease in intercepts, may, thus, reflect sample-level increase/decrease of latent traits (latent trait level) or manifest test-performance (indicator level). In turn, non-MI of intercepts can be interpreted similarly to increases/decreases in composite scores: It indicates changes of ability (on a latent factor level) or

test-performance (on the indicator level). Thus, this kind of invariance violation would not be a problem in longitudinal MI research, but reflects an anticipated effect.

Variances

Another indicator for performance change is variances, as these may (inter alia) increase if at least two groups of individuals develop in different directions. In contrast, whole-population changes in one direction would only result in intercept change, but not variance changes. Therefore, non-MI of variances (on latent and indicator levels) would not be a problem but could indicate subpopulations within the sample.

From this perspective, an increase in latent factor score variance may highlight that some participants depict no change in the target construct or even increases while others suffered from decreases. On the other hand, decreasing variances over time may indicate retest effects that diminish inter-individual differences in performance capability or the diminishing influence of variance-inducing third variables, such as trait anxiety (e.g., habituation effects), or simply normative aging processes that diminish smaller inter-individual differences over time.

However, other mechanisms may lead to similar changes in variance. For instance, increases in variance may also be attributable to increasingly fluctuating cognitive capabilities following cognitive decline and aging in general. Nonetheless, in the context of neuropsychological longitudinal MI research, non-invariance of variances may indicate that a certain domain is especially potent to distinguish healthy from abnormal courses or to at least indicate a certain cognitive domain to show some kind of aging-dependent variability.

Loadings

Another parameter that may show non-MI is the correlation of indicators and latent factors, which resembles weights within the composite approach. If loadings that previously were small enough to be neglected in increase to the extent that a new indicator should be added to the model or shifted from one latent factor to another, then the factor structure may change in its entirety (Cheung & Rensvold, 2002; Oort, 2005).

In the context of neuropsychiatric measures, the neuronal bases of performance in psychometric tests may change (e.g., verbal skill deficits may affect memory performance and lead to a reorganization of the factor structure). However, this effect may also be observed in normative age-related processes.

Nonetheless, regardless of the etiology of the loading shifts, invariance across measurement occasions would be a requirement of the classical composite approach as it implicitly assumes that all included variables contribute equally to the neuropsychiatric domain. Usually, weighted composite calculations are more beneficial. As weights of all variables entering a composite should reflect the loading of indicators on the latent factors, non-MI over time would imply that weights should also vary over time. Thus, non-MI is a general issue in this context and may highlight the shortcomings of classical composite approaches.

Longitudinal MI research based on neuropsychiatric test batteries

In contrast to the vast number of longitudinal research articles implemented on the prediction and the course of MCI/AD, far fewer of these have focused on latent factor structures and factorial invariance underlying cognitive domains within neuropsychiatric

test batteries to ensure generalizability (National Institute of Mental Health, 2011; Wicherts, 2016). Rather, some studies used the SEM approach to investigate between-group MI (Avila et al., 2020; Mitchell et al., 2012; Mungas et al., 2011; Sayegh & Knight, 2014; Tuokko et al., 2009). Others investigated latent factors and tested for MI in neuropsychiatric test batteries without keeping the longitudinal aspect in mind (Ma et al., 2021).

To our knowledge, only a few longitudinal measurement invariance studies, including the within-group latent factor approach based on neuropsychiatric test batteries, were published: For example, in a large multi-center sample of $N = 12020$ cognitively healthy participants and participants with diagnosed MCI or dementia were involved (age: ≥55 years; $M = 75.6$ years); researchers derived a four-factor structure from a neuropsychiatric battery (12 test variables), including the factor memory, attention, executive function, and language (Hayden et al., 2011, 2014). These factors remained invariant across the span of 1 year and predicted sample subgroups and cognitive impairment 3 years later. Moreover, Moreira et al. (2018) examined a two-factor model including memory performance and executive functioning in an elderly sample of 86 participants from a neuropsychiatric test battery. Defined factors remained invariant for 2 years. Similar studies concentrated on the two factors, namely, memory and executive functioning extracted out of large test batteries over periods of up to 8 years (Bertola et al., 2021; Williams, Chandola, & Pendleton, 2018).

Aims of the current study

As part of the prospective, observational, long-term, follow-up "Vogel Study" of a large German sample was conducted ($M = 73.9 \pm 1.55$ years of age at first out of three visits; see also Polak et al., 2017; Haberstumpf et al., 2020; Katzorke et al., 2018; Katzorke et al., 2017; Zeller et al., 2019); this current analysis aims to investigate longitudinal MI in a sample of (mostly) healthy elderly (at the first measurement occasion) over 3 years. However, in contrast to between-group MI-testing, we hypothesise and aim for the absence of MI, especially concerning variances of latent and manifest variables as these may indicate at least two groups of participants differing in their performance trajectory over time. Other mechanisms that may also result in increased variance may hint towards the importance of affected variables as potential targets for future studies. An Increased variance may result from the cognitive decline within the total sample (instead of within two distinct groups), which leads to more fluctuation in performance and, thus, longitudinal heteroscedasticity (Koscik et al., 2016). Nonetheless, non-MI would still provide for the insight that the affected variable is a valuable candidate for further investigation as it would have been indicated to be sensitive for cognitive decline or aging in general (see more on this in section 4). This non-MI may, thus, single out promising variables for further analyses as they possibly differentiate normal from pathological cognitive changes. Moreover, general decreases in intercept estimates (in both latent and manifest variables) are also anticipated, reflecting sample-based average changes in cognitive abilities on a latent level and changes in average test performance in manifest test scores. Additionally, MI of factor loadings is investigated to estimate possible shortcomings of the usual procedure to analyze sum-scores/composites.

Methods

Sample characterisation

As described earlier in Polak et al. (2017), the Vogel Study was carried out with the authorization of the local ethics committee (vote no. 23/11) and complied with the Helsinki Declaration (World Medical Association, 2013). Residents (with or without origin) of the city of Würzburg born between April 1936 and March 1941 (age: 70–77 years) were included in the study. All of them were informed about the project. They gave their written consent to participate in the Vogel Study, which started in the year 2011 and has now completed two out of three measurement time points (visit 1 [V1], visit 2 [V2], and visit [V3]). The project intends a total study duration of 10 years with 6 years of observation per participant.

Participants were excluded if they (1) suffered severe internal, psychiatric, or neurologic disease within the last 12 months (e.g., brain infarction) or (2) had a severe and uncorrected impairment of vision or hearing on the first day of data collection. Thus, a total of $N = 604$ subjects attended in the baseline examination of the Vogel Study.

At V2, approximately 3 years after V1, $n = 97$ participants no longer participated in the study ($n = 507$). This was, for example, due to death, the fulfillment of study exclusion criteria, study termination, relocation, or the deregistration of the telephone connection. For the current data analysis depicted below, participants who did not perform the neuropsychiatric test battery ($n = 125$) or exhibited more than five misses within the neuropsychiatric test battery ($n = 44$) because of rejection or high-stress experience at baseline or first follow-up examination were excluded. Even though this indicates dropouts to be dependent on personality or ability traits (e.g., cognitive abilities may have been worse in those who died within the next 3 years as existing disorders may have had impact at V1 already), we assume that the remaining misses within the final dataset were random.

We then calculated Mahalanobis-distances (cut off: $p < .001$; $n = 4$; Tab achnick & Fidell, 1996), as well as z-scores (cut off: ± 3.29; $n = 4$; Tab achnick & Fidell, 1996), for each neuropsychiatric test to find and subsequently exclude uni- and multi-variate outliers pairwise.

Therefore, the remaining sample of this article's final data set consisted of $n = 330$ participants (age: 70–77 years with $M = 73.78 \pm 1.52$ years at baseline examination; age: 73–81 years with $M = 77.67 \pm 1.60$ years at first follow-up examination; $n = 138$ females, $n = 192$ males; see Figure 1). So far, as described above, we still are in preparation for the second follow-up examination and have no data available yet.

Neuropsychiatric test battery

Besides the examination of various demographic, anamnestic (e.g., lifestyle, medical history, etc.), affectivity, autonomy, blood, and lifestyle variables to characterise our sample, we conducted a neuropsychiatric test battery comprising: (1) the Verbal Learning And Memory Test (VLMT; Helmstaedter, Lendt, & Lux, 2001), (2) the Wechsler Memory Scale-Revised (WMS-R; Härting et al., 2000), (3) the Regensburger Verbal Fluency Test (RWT; Aschenbrenner, Tucha, & Lange, 2000), (4) the Rey Complex Figure Test (CFT; Fimm & Zimmermann, 2001; Meyers & Meyers, 1996), and (5) the battery of Tests for Attentional Performance (TAP; Fimm & Zimmermann, 2001). For a more detailed description of the general examination procedure within the Vogel Study, see our previous method studies (Polak et al., 2017).

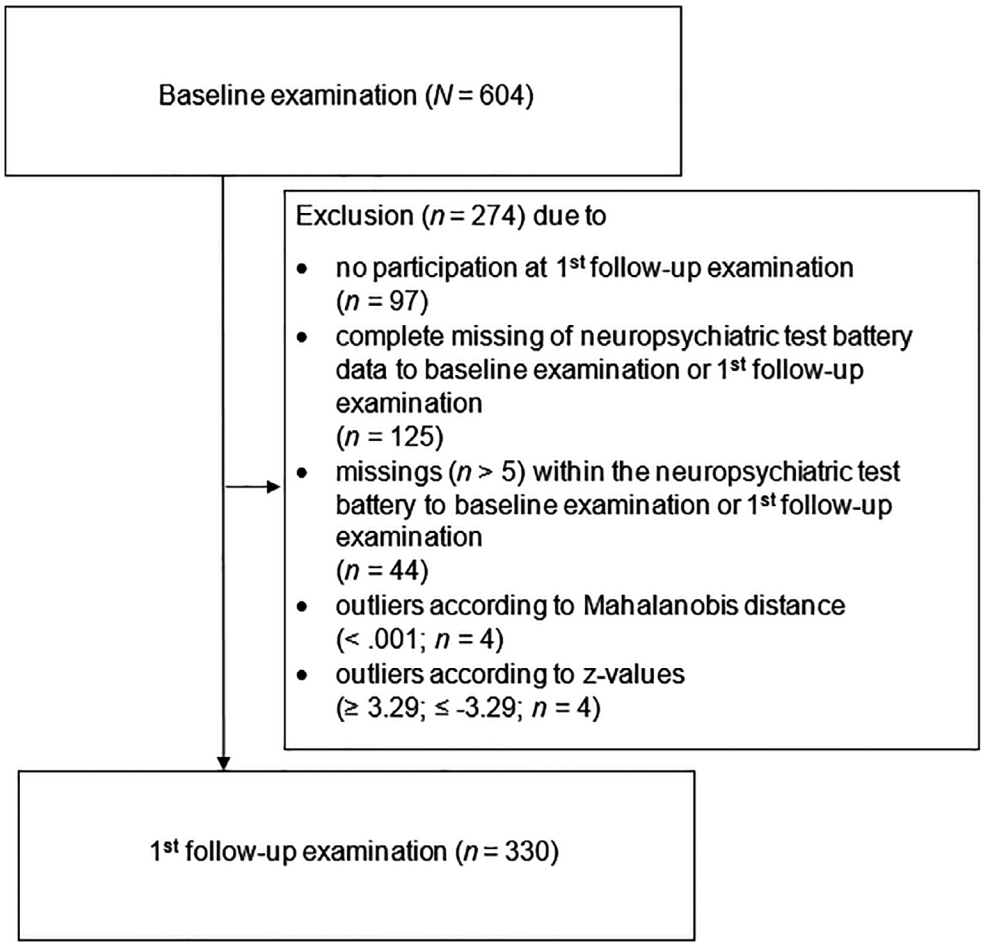

Figure 1. Course of exclusion for data analysis; CNS = central nervous system.

The subsequent test scores were used in further latent factor analyses: VLMT immediate recall (sum score words), VLMT delayed recall (sum score reproduced words), VLMT recognition (sum score recognition word list), WMS-R digit span (sum score), WMS-R block span (sum score), RWT verbal fluency (sum score), RWT category fluency (sum score), CFT memory (sum score both reproduction times), CFT visuoconstruction (drawing score), TAP tonic alertness (median of reaction time [RT]), TAP phasic alertness (RT-parameter for phasic alertness), TAP divided attention (omission error), TAP GoNoGo (error number), and TAP incompatibility (F-value of "field of vision x hand" interaction). Thus, the following latent factor analysis comprised 14 test variables detached from five neuropsychiatric tests.

Statistical analyses

The data preparation, outlier detection, testing of prerequisite assumptions, and the EFA were conducted in IBM SPSS Statistics for Windows (version 25; SPSS Inc). Further CFA analyses were completed in R (lavaan package version 0.6–5; (Rosseel, 2012; R Core

Team, 2016). Predictive mixed models were also fitted via R (lme4 and lmerTest packages; Bates, Maechler, Bolker, & Walker, 2014; Kuznetsova, Brockhoff, & Christensen, 2015, 2017).

Acceptable cut-offs for fit indices, for example, the root mean square error of approximation (RMSEA) and comparative fit index (CFI) were set to <0.05 and >0.95, respectively. The alpha level to test for significance in χ^2-tests was set to <0.05.

Regarding the SEM, standardizing manifest variables may lead to biased estimates in longitudinal data (Kline, 2005; Schumacker & Lomax, 2004). Also, some tests did not provide samples that qualified for T-value calculation in all ages of participants who were included in this study. To get an unbiased estimation of course effects, raw test scores were used for further latent factor analyses (13 raw scores and 1 *F*-value for TAP incompatibility[1]).

Moreover, as unstandardised test scores exhibited substantial differences in their respective scales, those tests depicting variances greater than 10 times the magnitude of the smallest variance found in the dataset were rescaled. This procedure is thought to diminish chances for Heywood cases and other estimation issues (Kline, 2005; Schumacker & Lomax, 2004). Finally, reaction time–based variables were transformed via natural logarithm (TAP tonic and phasic alertness). However, no other transformation was carried out, which led to non-normality of several test scores. Even though this may, in theory, impair reliable estimation, several simulation studies reported only a small non-normality impact on standard errors (Lei & Lomax, 2005) or model fit (Gao, Mokhtarian, & Johnston, 2008). Furthermore, since the effect of non-normality may vary across different estimation methods, robust maximum likelihood estimation was used. This function leads to reliable model estimations considering mis-specification, non-normality of data, and/or small sample sizes (Gao, Shi, & Maydeu-Olivares, 2020; Lai, 2018; Yilmaz, 2019).

Exploratory factor analysis (EFA)

To find a fitting latent factor structure, an EFA was carried out, including data of both measurement occasions. A parallel analysis was carried out to define the number of factors that were subsequently extracted after Varimax rotation. The Kaiser–Meyer–Olkin (KMO) criterium and Bartlett's test of sphericity were assessed to ensure suitable prerequisites for the analysis. Only those tests depicting rotated loadings of four or higher on only one factor were included in the final model.

Invariance testing

The concluding factor structure, indicated by the EFA, was tested in a multi-group CFA using full information maximum likelihood estimation in the handling of misses, the lavaan-default "nlminb" optimization method, and robust maximum likelihood estimation (MLR) for the calculation of standard errors. Groups were defined by test sessions, which were 3 years apart, enabling a longitudinal interpretation of cross-group effects. Each participant remaining in the dataset was present on both occasions.

As stated before, MI is usually tested via increasingly restrictive CFAs. In this context, "restriction" refers to the fact that a given parameter is not allowed to vary across groups

[1] *This test calculates an F-test to evaluate slowing in reaction time due to incompatibility compared to compatible trials in a flanker task (the higher the percentage rank, the lower the incompatibility effect).*

(measurement occasions): Suppose the fit between a predefined model and the actual data decreases by imposing such a restriction, in that case, this restriction seems to have violated the actual data structure in the sense that the data would be better represented by allowing varying parameters across groups, indicating non-MI.

Hence, each of the following models adds certain parameters to the previous models' restrictions. Comparing model fit across these, significant decreases in fit indices would indicate non-invariance (the restricted parameter varies over time). To test this, χ^2 statistics were calculated. These statistics indicate differences between one model and the model before (model 2 vs. 1, model 3 vs. 2, and model 4 vs. 3). Following theoretical remarks, a total of four models were fit (Cheung & Rensvold, 2002; Dowling, Hermann, La Rue, & Sager, 2010; Van de Schoot, Lugtig, & Hox, 2012), including the following:

Configural model. In this model, only the factor structure (assignment of tests to latent factors) implied by the EFA was restricted for all variables. Otherwise, this model is built to freely estimate as many parameters as possible. However, to ensure the model to be identifiable, some restrictions need to be made. In this study, two separate approaches are discussed to give examples on possible modeling decisions concerning two different-use cases.

First, to investigate measurement invariance with a focus on manifest-latent- factor-interaction, the loading of one indicator variable per factor was restricted to 1. Also, the means/intercepts of the latent factors were restricted to 0 to give the latent factors a metric. Since means of the latent factors are not allowed to differ from 0, changes within latent abilities will be propagated to manifest test score intercept differences over time, enabling the investigation of test properties (i.e., how well they are suited to investigate latent ability changes). This approach was used at first.

In addition, one may consider the extraction of latent ability scores for further investigation (e.g., to use it as dependent variables within regression analyses or ANOVAs). Thus, for this goal, it is more beneficial to allow free latent score estimation at the second measurement occasion. To do so, in an exemplary use case, the configural model was later refitted with a restriction of latent variable means to 0 and latent variable variances to 1 for the first measurement occasion only. Furthermore, loadings of one manifest indicator variable per factor were restricted to be equal across both measurement occasions, which enabled the model to estimate latent factor means and variance freely at the second measurement. Thus, in this model, significance of changes over time can be easily assessed by investigation of latent variable estimates at V2 (intercepts are significant if they differ significantly from 0, variances are significant if they differ significantly from 1).

Regardless of these modeling choices, overall (absolute) fit of this kind of model indicates that the model structure (association of tests to a certain latent factor) is invariant over time. If this was violated, latent abilities would not be indicated by the same tests across time, which would imply severe issues with the composite approach and question the validity of course data in general.

Metric model. In the next model, investigating (construct-level) metric invariance, all loadings across groups/time were restricted to equal one another. The means of the factors themselves were still fixed to 0, while the loading of one indicator per factor was fixed to 1. In this model, invariance implies that manifest test scores equally indicate the

given latent constructs over time. Violation of this loading invariance would imply that the weights of variables used for composite approaches must be adjusted over time.

Scalar model. The third, scalar model, added a cross-group restriction of manifest indicator intercepts. By doing so, the measurement model is identifiable without latent mean fixation. Thus, latent means were estimated freely instead of being fixed to 0. In this model, non-MI across groups indicate changes in the difficulty of tests (changes in performance by participants). Furthermore, latent factor intercepts may be analyzed to find longitudinal decreases/increases in latent abilities. Violation of intercept invariance would not pose a problem but may indicate anticipated effects of ability/performance decline.

Variance model. Finally, in addition to these restrictions, variances of latent factors were held constant across groups/time. Non-invariance in this model may reflect the presence of at least two groups of participants whose latent abilities evolve into different directions over time or the presence of other mechanisms that affect the overall variability of measured ability within the whole sample. Thus, violation of the invariance assumption would be in line with anticipated effects as this may highlight variables/parameters that could possibly be best suited for detection of early MCI-related whole sample or subsample–based changes (e.g., healthy vs. abnormal cognitive courses).

Composite approach

To assess the benefit of latent-factor-score analysis with the more common composite approach, unweighted composite variables were calculated for comparison. To do so, the test score of each subject was standardised for each individual test by placing the score obtained in the context of an age- as well as gender- and education-matched norm sample (all test scores except the VLMT and CFT). In total, four composites were calculated before the context of the factor structure defined by the EFA by simply averaging test scores assigned to a common factor (see Figure 3). The models investigated the same $n = 330$ participants.

To then compare the benefit of the latent factor approach over the unweighted composites, as an example, a mixed effect regression model was fit once with the latent factor estimate for declarative memory as a dependent variable and once with the respective composite as such. As a result, the two models can be compared directly by comparing the estimated effects of predictors (which are the same across both models) for these two dependent variables.

Results

Exploratory factor analysis

Both the KMO criterium (.688) and Bartlett's test of sphericity ($\chi^2(91) = 1974.583$, df = 91, $p < .001$) implicated suitable prerequisites to conduct the analysis. Subsequently, a total of five factors were extracted following the suggestions of both Eigenvalue and parallel analysis. Estimations of factor properties and a scree plot are shown in Table 1 and Figure 2. Rotated loadings ≥ 0.4 are displayed in Table 2.

Table 1. Estimations of factor properties

	Eigenvalue	Explained variance	Cumulative explained variance
Factor 1	2.463	17.591	17.591
Factor 2	2.007	14.339	31.930
Factor 3	1.827	13.052	44.982
Factor 4	1.705	12.176	57.158
Factor 5	1.091	7.792	64.950

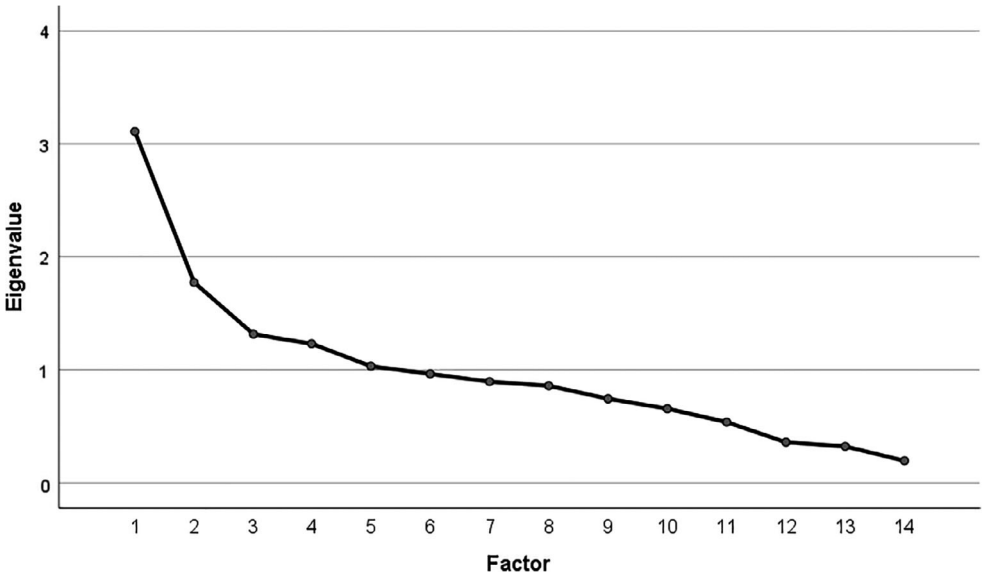

Figure 2. Scree plot showing the five-factor solution of the Exploratory Factor Analysis (EFA).

Cognitive domains were assigned to describe the factors as denominated in Table 3. However, only four factors of those implicated by the EFA were analyzed further as the fifth factor comprised only one indicator complicating estimation (Kline, 2005; Schumacker & Lomax, 2004).

Measurement invariance testing

Four increasingly restricted models were fit and compared to analyze measurement invariance (see Table 4). Both the RMSEA and CFI indicated acceptable model data assuming that the assignment of manifest test scores to latent factors stays equal across time. Hence, the conceptual representation shown in Figure 3 represents the suitable structure for both measurement occasions. However, Table 4 further summarises that factor loadings, test intercepts, latent means, and latent variances depict substantial non-MI over time.

However, as non-MI is not per se a property of the whole model, but rather of certain parameters, further analyses were carried out to clarify which test- parameters significantly changed over time and which did not. To assess this, the configural model

Table 2. Factor rotation of the five-factor solution of the exploratory factor analysis (EFA)

	Factor loadings after varimax rotation				
Scale	1	2	3	4	5
VLMT immediate recall	0.898	–	–	–	–
VLMT delayed recall	0.888	–	–	–	–
VLMT recognition	0.861	–	–	–	–
TAP tonic alertness	–	0.997	–	–	–
TAP phasic alertness	–	0.997	–	–	–
WMS-R digit span	–	–	0.536	–	–
RWT verbal fluency	–	–	0.833	–	–
RWT category fluency	–	–	0.867	–	–
WMS-R block span	–	–	–	0.565	–
CFT memory	–	–	–	0.724	–
CFT visuoconstruction	–	–	–	0.744	–
TAP compatible	–	–	–	–	0.888
TAP divided attention	–	–	–	–	–
TAP GoNoGo	–	–	–	–	–

EFA coefficients ≥0.40 are exhibited. VLMT = verbal learning and memory test (Helmstaedter et al., 2001); TAP = battery of tests for attentional performance (Fimm & Zimmermann, 2001); WMS-R = Wechsler Memory Scale-Revised (Härting et al., 2000); RWT = Regensburger verbal fluency test (Aschenbrenner et al., 2000); CFT = Rey complex figure test (Meyers & Meyers, 1996).

Table 3. Designation of the four latent factors

Latent Factors	Cognitive Domain	Included neuropsychiatric test scores
Factor 1	declarative memory	VLMT immediate recall, VLMT delayed recall, VLMT recognition
Factor 2	attention	TAP tonic alertness, TAP phasic alertness
Factor 3	working memory	RWT verbal fluency, RWT category fluency, WMS-R digit span
Factor 4	visual-spatial processing	CFT memory, CFT visuoconstruction, WMS-R block span

VLMT = verbal learning and memory test (Helmstaedter et al., 2001); TAP = battery of tests for attentional performance (Fimm & Zimmermann, 2001); RWT = Regensburger verbal fluency test (Aschenbrenner et al., 2000); WMS-R = Wechsler Memory Scale-Revised (Härting et al., 2000); CFT = Rey complex figure test (Meyers & Meyers, 1996).

Table 4. Confirmatory Factor Analyses (CFAs) for the sample of $n = 330$ participants. Reported fit-parameter base on a robust maximum likelihood estimation

CFA model	RMSEA	CFI	AIC	BIC	χ^2	df	*p*
Fixed structure	0.049	.969	22068	22418	135.22	76	
+ Fixed loadings	0.051	.963	22073	22392	154.33	83	.007**
+ Fixed intercepts	0.080	.902	22182	22469	277.16	90	<.001
+ Fixed variances	0.097	.849	22281	22550	384.20	94	<.001

RMSEA = root mean square error of approximation; CFI = comparative fit index; AIC = Akaike information criterion; BIC = Bayesian information criterion.

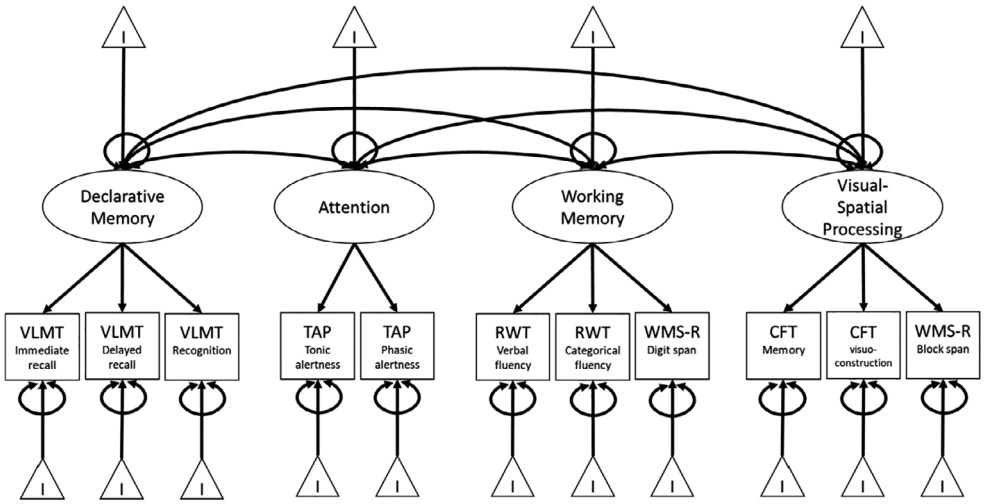

Figure 3. A conceptual model for the CFAs depicting estimated parameters. Oval variables depict latent factors, while rectangles reflect manifest test scores. Triangles represent intercepts.

yields the best insight as all the more restrictive models showed decreasing fit to the data. Furthermore, as the configural model allows for the greatest number of freely estimated parameters, non-invariance in the data that would influence the fit indices of more restrictive models negatively should be reflected in significantly changing estimates over time. Table 5 summarises the results by providing estimates across both measurement occasions. Furthermore, to provide information on the tendency of statistical significance of descriptive differences, the standard error of a variable's mean was multiplied by 1.96. By doing so, a 95% confidence interval (CI) was obtained. If the CI of either value (measurement occasion 1 or 2) included the estimated mean of the other measurement occasion, no significant difference was assumed. Please note that this comparison included two tests for each variable, which were not corrected for. Shading in Table 5, thus, indicates trends (exploratory findings), but not confirmatory hypothesis testing, as no specific assumption on non-MI of specific parameters was formulated beforehand.

Nonetheless, results indicate partial non-MI for loadings in VLMT recognition and immediate recall (declarative memory) and WMS-R block span (visual–spatial processing). Furthermore, VLMT immediate recall, VLMT recognition, and CFT visuoconstruction intercepts seem to decrease while CFT memory, tonic alertness, and WMS-R scores increase. Finally, VLMT delayed recall, RWT category fluency, and WMS-R digit span scores also seem to decrease in their variance over time, while two of the three working memory–related scores RWT verbal fluency and CFT visuoconstruction increase in variance.

Furthermore, in addition to the (manifest) indicator-level analyses of Table 5, latent factor estimates are summarised in Table 6. While indicator-level estimations of Table 5 were made following the restriction of one indicator variable loading per factor to 1 and latent means to 0, the results in Table 6 were obtained by restricting the loading of one indicator per factor to the same value across groups while setting the latent means to 0 and the latent factor variance to 1 for the first measurement occasion, allowing for free estimation of these parameters at the second occasion. By doing this, free estimation of latent factor parameters could be ensured, which would be necessary for subsequent longitudinal prediction analyses using these latent factor scores as dependent variables.

Table 5. Model parameters for a configural model

	Unstandardised loadings		Intercepts		Variances	
Variables and factors	V1	V2	V1	V2	V1	V2
VLMT delayed recall (F1)	1	1	6.004 (0.119)	5.437 (0.183)	2.461 (0.383)	0.575 (0.193)
VLMT recognition (F1)	0.731 (0.032)	0.410 (0.018)	5.391 (0.104)	4.858 (0.130)	1.692 (0.192)	1.913 (0.247)
VLMT immediate recall (F1)	0.810 (0.031)	0.620 (0.027)	9.704 (0.098)	8.990 (0.171)	0.829 (0.119)	1.262 (0.212)
TAP phasic alertness (F2)	1	1	23.128 (0.067)	23.172 (0.109)	0.012 (0.003)	3.423 (1.402)
TAP tonic alertness (F2)	1.754 (0.718)	2.840 (1.198)	26.653 (0.114)	26.974 (0.123)	0.036 (0.005)	0.979 (1.836)
RWT verbal fluency (F3)	1	1	2.701 (0.068)	2.672 (0.062)	0.286 (0.111)	0.534 (0.081)
RWT category fluency (F3)	1.090 (0.146)	1.125 (0.117)	3.534 (0.061)	3.533 (0.061)	0.724 (0.122)	0.296 (0.089)
WMS-R digit span (F3)	0.799 (0.145)	0.829 (0.138)	8.291 (0.117)	9.558 (0.107)	4.004 (0.276)	3.263 (0.277)
CFT visuoconstruction (F4)	1	1	34.836 (0.107)	34.102 (0.160)	2.668 (0.375)	4.216 (0.793)
CFT memory (F4)	1.389 (0.357)	1.368 (0.289)	7.259 (0.144)	7.811 (0.165)	4.62 (0.617)	2.825 (1.157)
WMS-R block span (F4)	0.490 (0.112)	0.217 (0.050)	7.394 (0.078)	6.982 (0.079)	1.743 (0.146)	1.894 (0.157)

V1 = Visit 1, V2 = Visit 2. Unstandardised estimates and standard errors (SEs; in parentheses) are reported. Light grey cell shadings reveal significant increases of estimates over time, dark grey cell shadings reveal significant decreases, which indicates non-MI.

Table 6. Latent factor model parameters for a configural model

	Covariances		Intercepts		Variances	
	V1	V2	V1	V2	V1	V2
Declarative memory						
Attention	−0.004 (0.054)	0.089 (0.136)	0	−0.568 (0.218)	1	9.737 (1.196)
Working memory	0.355 (0.052)	0.718 (0.181)				
Visual–spatial processing	0.364 (0.070)	1.716 (0.369)				
Attention[a]						
Working memory	−0.244 (0.089)	−0.187 (0.093)	0	0.044 (0.128)	1	0.499 (0.271)
Visual–spatial processing	−0.179 (0.104)	−0.169 (0.096)				
Visual-spatial processing						
Working memory	0.343 (0.094)	0.478 (0.151)	0	−0.734 (0.193)	1	3.283 (0.982)
Working memory	–	–	0	−0.029 (0.092)	1	0.733 (0.107)

V1 = Visit 1, V2 = Visit 2. Unstandardised estimates and standard errors (SEs; in parentheses) of latent factor estimates are reported. Light grey cell shadings reveal significant increases of estimates over time and dark grey cell shadings reveal significant decreases, which indicates non-MI. These results indicate increases in the co-dependency of declarative memory and visual–spatial processing over time. Furthermore, it seems that the latent ability of declarative memory as well as visual–spatial processing decreased on average over time. Finally, the variance of declarative memory, working memory, and visual–spatial processing increased as well.

[a]This factor is estimated by variables expressing reaction times. Thus, higher values indicate worse performance.

To demonstrate this idea, this model was used for parameter extraction as it imposes the least restrictions while allowing free latent factor estimation.

Before the findings reported above, models fixating more parameters fit the data significantly worse. Hence, again, this model provides the most unbiased estimates. Table 6 highlights that the covariance between declarative memory and visual–spatial processing increases over time. Furthermore, intercepts decrease in declarative memory as well as in working memory. Finally, variances increased in both declarative memory and visual–spatial processing.

Comparison between the latent factor approach and the composite approach

Figure 4 illustrates the course of latent factor means and composites across V1–V2. Descriptively, both the composite and latent factor approach indicate decreasing performance/ability scores for declarative memory over time. However, the latent factor approach indicates greater significance and effect size. Regarding attention, again, both approaches indicate a similar trend, this time towards increases in scores. Since attention scores are indexed by reaction time, depicted increases indicate decreases in reaction speed (thus, worse performance/capability) with a seemingly greater effect estimate in the composite approach. In working memory, the composite approach suggests increases in performance over time, while the latent factor approach shows no particular change. Finally, concerning visual–spatial processing, a significantly greater decrease in scores is shown in the latent factor than the composite approach over time (indicating a decrease in processing capability).

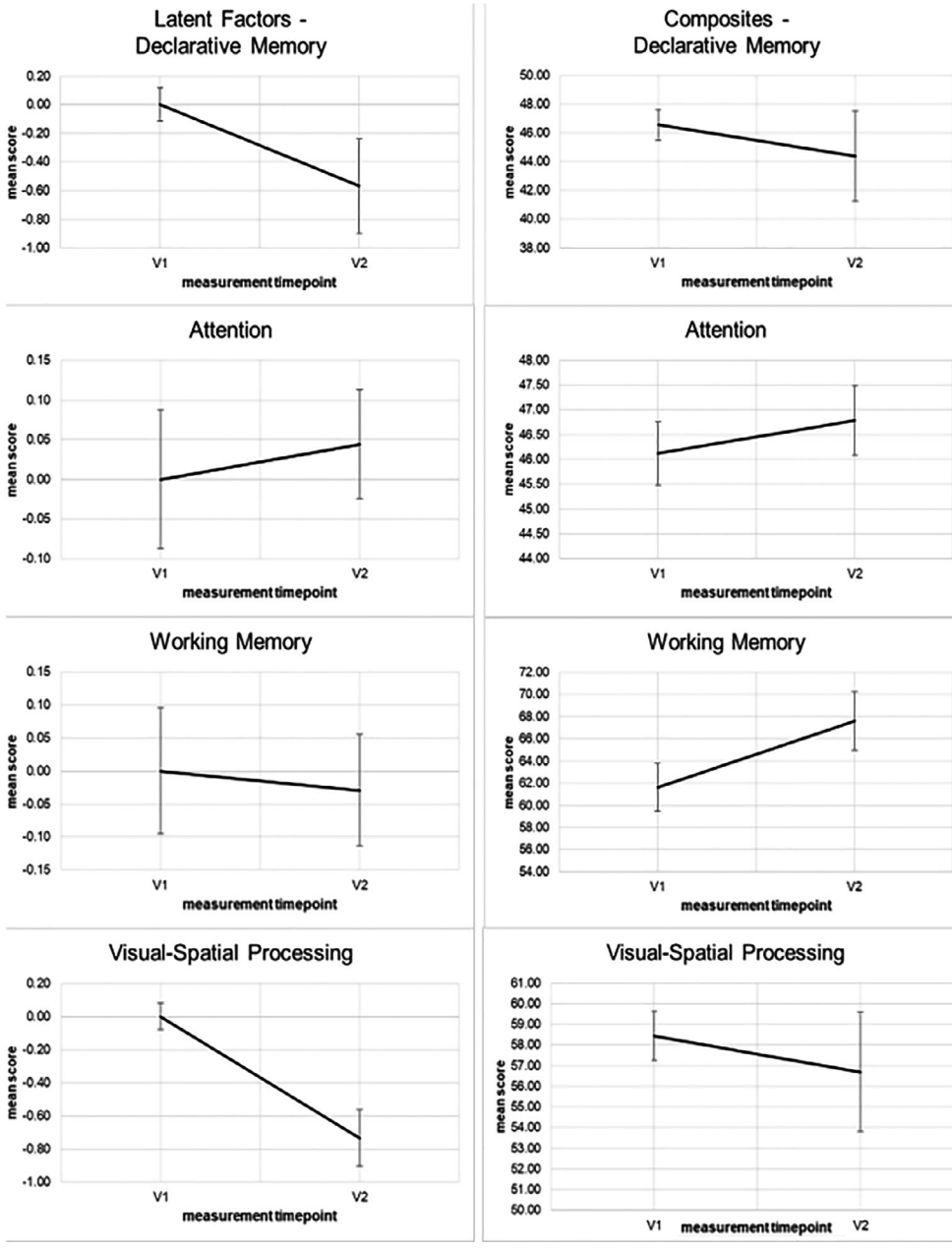

Figure 4. Latent factor approach (left column) compared with the composite approach (right column). V1 = Visit 1 and V2 = Visit 2. Lines indicate mean (M) and error indicators represent 95% confidence interval (CI).

Example procedures for the prediction of latent ability scores

The current article focuses on the general applicability of the SEM approach for describing the course of cognitive abilities and their decline. While the central part of this article is based on the description of measurement invariance as a potential source of information

for such questions, the estimated latent ability scores of the analysis can also be used directly to test specific predictor variables for their predictive ability. Two ways of doing this will be briefly described:

First, the predictor variables themselves can be included in the model. The SEM approach allows both latent and manifest scores to be predicted by both fixed (e.g., genetic vulnerabilities) and variable predictors (e.g., depression scores varying over each measurement occasion). Thus, for datasets with at least three measurement occasions (for linear trends; more may be needed for non-linear trajectories; e.g., Byrne & Crombie, 2003; Felt, Depaoli, & Tiemensma, 2017; Grimm & Ram, 2009), a latent growth curve model could be defined, in which the second-order latent factors are assumed that define a slope across the measurement occasions as well as an intercept, influencing the first-order latent ability factors. At the same time, if one assumes that not all latent abilities (e.g., memory vs. visual–spatial processing) show the same slopes over time, a second-order latent slope and intercept could also be defined for each latent ability factor individually. These factors (slope and intercept) can, in turn, be predicted by predictor variables, making it possible to predict the temporal variation of latent factor scores with parameters such as genetic vulnerability factors. At the same time, the manifest variables measured per time point can also be predicted by influences that also change over time (e.g., BDNF levels, depression scores). This modeling approach establishes a link between measured values that would, otherwise, be mistakenly treated as between-effects rather than within-effects (e.g., the manifest test scores of a person in a test at two measurement times). However, such estimation would require significantly more study participants (Willett & Sayer, 1994), with the advantages of directly estimating the influence of predictors at the manifest and latent levels while simultaneously allowing for measurement invariance estimation.

As a second possibility, the latent factor scores could be read from the model and included as a dependent variable in a regression analysis or ANOVA. To illustrate this approach (which is also possible with the current dataset), a mixed model approach was chosen in which various predictor variables collected in the study were used as independent variables. In contrast, the extracted latent factor scores were used as dependent variables.

The dependent latent factors were included with a random intercept for each subject and time as a fixed effect predictor (levels *one* and *two*, model 1). Due to their relevance in the literature (for an overview, see Xu et al., 2015), and supposedly low multicollinearity (or redundancy), the covariates age and gender (model 2), Brain-Derived Neurotrophic Factor (BDNF; quantified in serum by ELISA; model 3), depressiveness (measured by the Beck Depression Inventory-II [BDI-II]; Beck, Steer, & Brown, 1996; model 4), and vitamin B12 (quantified by blood plasma; model 5) were included as potential predictors (see Appendix A for more details). Overall, one set of models was tested for each latent factor. The model with the highest fit index was subsequently chosen as the best model for the interpretation of fixed effects. Results revealed plausible predictive effects, mostly involving gender, age, and their respective interaction with time for all latent factors, except for attention. For instance, concerning declarative memory, the best model revealed a significant main effect for age ($\beta = -0.11$, $t\,(299) = -2.340$, $p = .019$), time ($\beta = -0.13$, $t\,(299) = -2.898$, $p = .004$), and gender ($\beta = 0.44$, $t\,(299) = 4.765$, $p < .001$), indicating declining scores with higher age and over time in men as compared to women.

To draw a comparison between the classical composite approach and the latent factor approach presented here, the same model was again fitted with an unweighted composite

approach. Thus, the same predictors and their respective interactions were used with the composite *declarative memory* as the dependent variable (see section 2.3.3. for details). This mixed model revealed only one effect for gender ($\beta = 0.22$, t (299) = 2.483, $p = .014$), which was also less significant as compared to the latent factor approach. As a result, concerning the declarative memory domain, the same predictor model within the latent factor approach was able to find more significant and greater effects for the given list of predictors as than the composite approach.

Discussion

The current longitudinal analysis was performed to identify hints towards cognitive decline in a sample cohort totaling $n = 330$ individuals. As part of this, longitudinal MI of a test battery of 14 neuropsychiatric test variables was investigated across 3 years, which led to the identification of four stable latent factors of cognitive abilities: declarative memory, attention, working memory, and visual–spatial processing. Furthermore, predictive analyses using scores of these domains as a dependent variable indicated that latent ability scores increased significance of regression analyses in comparison to composite scores.

Longitudinal measurement invariance

To date, there are only few studies available that analyzed longitudinal MI in neuropsychiatric test batteries and defined latent factor scores as dependent variables in the prediction of pathological cognitive decline. The SEM approach allows for concurrent testing of group-/time-related differences in latent and manifest variables. Table 4 indicates that loadings of indicator variables significantly vary over time, leading to small but significant changes in model fit. Table 5 further clarifies this non-MI which stems from VLMT-, RWT-, and WMS-R–related measures. In the context of the classical composite approach, this highlights a possible reason for null findings: In the current study, the WMS-R block span test score becomes less indicative of the latent visual–spatial processing performance. Thus, if a researcher imposes the same weighting on this test score at both measurement occasions, the resulting trend over time may be biased.

To put the study results in a simple context, the following section compares approaches by using loadings of the configural model as weights while using actual mean standardised test scores (standardised to $M = 50$, SD = 10) of the $n = 330$ participants as variables. For unweighted composites, researchers may usually use the formula shown in equation (1):

$$(1 \times 73.95 + 1 \times 59.76 + 1 \times 41.74) - (1 \times 70.20 + 1 \times 66.0 + 1 \times 35.07) = 4.18 \quad (1)$$

Here, (1) indicates the implicitly imposed weight of 1 for each variable and other numbers reflecting the average test score. By subtracting the scores of the second measurement occasion from the first, the result reflects the mean change of the visual–spatial processing composite over time. A positive score indicates decreases over time. However, Table 5 suggests that performances in the three tests are not equally relevant to the latent ability of visual–spatial processing. Thus, equation (2) modifies the previous approach by imposing different weights for each variable (according to Table 5):

$$(1 \times 73.95 + 1.389 \times 59.76 + 0.490 \times 41.74) - (1 \times 70.20 + 1.389 \times 66.0 + 0.490 \times 35.07) = -1.65 \quad (2)$$

The result indicates an overestimation of the unweighted composite approach in comparison to the weighted one. The latter implies an overall increase in skill. However, equation (2) would be appropriate only if measurement invariance for factor loadings was given. Since Table 3 indicates otherwise, the formula is once again adapted by varying weights over time. The resulting equation is given in equation (3):

$$(1 \times 73.95 + 1.389 \times 59.76 + 0.490 \times 41.74) - (1 \times 70.20 + 1.368 \times 66.0 + 0.217 \times 35.07) = 9.31 \quad (3)$$

Since equation (3) results from the model that fits the data best, we assume that its result is the most unbiased. Equation (1) would only be unbiased if all test scores were equally relevant/indicative/correlated to/of the target construct (visual–spatial processing), while equation (2) would only hold if metric invariance was given. This example illustrates the value of SEM-driven course analyses and possible shortcomings of the most often used approach as equation (3) produced an effect more than double in size as the unweighted composite approach did.

The maximal mis-estimation due to non-MI is given by the sum of absolute factor loading differences between measurement occasions (Schmitt, Golubovich, & Leong, 2011). For instance, in this study, the sum of the factor of visual–spatial processing equals 0.137 (| (0.410–0.487) |+| (0.620–0.540) |.

Furthermore, apart from changes in loadings over time, test intercepts also partially varied. For instance, the WMS-digit span test intercepts significantly increased while WMS-block span performance decreased (see Table 5). This result highlights that even though these two subscales were taken from one test battery, the performance trajectories were opposed to one another. Interestingly, this fits well with the factorial structure. While both visual–spatial processing and declarative memory seem to pose as promising cognitive domains to assess early changes in abilities (indicated by Table 6), working memory was mostly characterised by invariance over time, which may be why mostly VLMT measures, along with WMS-block span showed decreasing intercepts in Table 5.

Course of latent factor scores

On a latent factor level, variances of declarative and working memory, as well as visual–spatial processing, increased over time, indicating the existence of at least one mechanism that may drive increases in sample-based variance statistics of latent ability scores. One explanation may lie in the existence of at least two groups that develop in different directions. At the same time, another possibility is given by the increasing inter-individual variance in "ability retrieval" in those who suffer from cognitive decline (memory capacitance may vary more greatly from day-to-day in those who show signs of an MCI than it does in healthy young adults). Either way, this indicates that declarative memory, working memory, and visual–spatial processing pose as early indicators for age-related changes in cognition.

Moreover, in the context of neuropsychiatric test scores, differences in covariances, including factor loadings and latent factor covariance, may reflect compensatory mechanisms between initially independent neuronal systems and functions. On that

note, the connection between visual–spatial processing and declarative memory increased over time, possibly indicating that at least one subsystem relies on the other increasingly. Other interpretations may assume a third variable to produce these changes in covariance matrices. For instance, an uncontrolled third variable may affect both factors, thereby increasing their correlation over time. One such factor may be early signs of cognitive decline, which would be plausible given that only test intercepts of these two factors expressed non-MI. However, these interpretations are not yet reliable, basing them solely on the data of this one study. Further research and discussion are needed. Nonetheless, since covariance between those two factors that show the greatest changes in variance and intercepts increased, it would be plausible to assume that the connection between these constructs increases as a function of age and/or pathology-related cognitive decline.

Prediction of pathological cognitive decline

Finally, the current study provided a short example on the topic of latent score prediction based on covariates, psychometric parameters, and biomarkers.

Since the exemplary effects in latent factor score–related analyses were generally more significant than in the composite models, these data highlight the possible benefit from investing in the more complex but possibly more reliable and valid SEM approach. This is especially relevant since the composite models produced effects that may not be plausible, such as increased working memory capability in older individuals.

Conclusions

Methodologically, it may be appropriate to calculate (weighted) composite variables instead of latent factors. However, it is important to note that composite variables do not adapt to the data over time. This indicates active neglect of compensation mechanisms, retest effects, habituation to test settings, the influence of increasingly severe diseases, and many more factors of influence as these may cause significant changes in the interdependence ability of neuropsychiatric functioning. For instance, loss of function in certain brain areas may affect the inter-correlation of neuropsychiatric domains by making them dependent on other compensating areas/functions. Moreover, psychological variables such as trait anxiety may impair performance in the first measurement occasion to other extents as it does in the second due to habituation effects.

SEM, on the other hand, estimates such influences indirectly by addressing changes in the correlation matrix among manifest test scores obtained. As a result, changes in manifest scores, latent performances, and their correlation can be addressed all at once. In fact, in this study, we found hints to either compensate mechanism or neurological change over time as the covariance between visual–spatial processing and declarative memory increased. Since the variance of both latent factors also increased, while their intercepts decreased, these results may hint at least two sub-samples within the analyzed participants that showed different trajectories in their cognitive abilities across measurement occasions or a general decline in capabilities on these domains that results in increased inter-individual variability of skills (or a mix of both). However, this interpretation is speculative and needs clarification by identifying predictors for these latent score changes. The above-mentioned interpretation would become very plausible if the covariance pattern between both factors would decline to their baseline level after

controlling for such predictors. Hence, again the SEM approach provides additional ways to gain more detailed insights into the data as the composite approach does. Moreover, in the example for one of the latent factors provided above, we were able to show that the composite approach underestimated the effect of change over time by more than 50%, which again highlights possible shortcomings of the classical composite approach and may reveal mechanisms by which classical longitudinal analyses may have trouble finding reliable and significant change.

Ultimately, the exemplary prediction analysis depicted in this study provided further evidence of the superiority of the SEM approach over the composite approach, supporting the idea that this approach produces more reliable results.

To conclude our findings, this study was able to find four latent factors that are in line with the previous research. Furthermore, by testing these factors for longitudinal measurement invariance, this study provides insights into calculating the extent of bias that may lead to inflation or false null findings in the classical composite approach. In addition, even though measurement invariance was not present for most parameters, this study also discussed how this may be beneficial in understanding both normative and pathological aging. In summary, the SEM approach adds highly relevant information to the interpretation of longitudinal neuropsychiatric data.

Limitations

First, the generalisability of the factor structure may be impaired as the results are specific to the neuropsychiatric test battery used. Also, although supported by the residents' registration office, participant recruitment was not fully representative for the general population (Polak et al., 2017). In the case of the Vogel Study, participants had a relatively higher education level in than the general German population (Statistisches Bundesamt, 2018).

Additionally, one problem of this latent factor model was the factor attention as it comprised only two manifest indicators, which may have significantly biased results for this domain (Kline, 2005). This may be one reason why no significant effect was present regarding this factor. Nonetheless, we do not anticipate significant mis-estimation of other factors and their indicators due to this issue.

Also, we want to note that the use of factor scores may propagate estimation errors within the SEM to the analysis of predictor variables which is an inherent risk to this approach and may lead to false results.

Due to relatively small sample size concerning the complex methodology, the precision of model estimation may have suffered. However, larger sample sizes may lead to smaller error terms and increased significance even for small non-invariance, which may also pose an issue as this may lead to overly sensitive analyses. As a result, we argue to estimate the difference of effects due to variability of parameters (by using the formulas (1), (2) and (3)) to estimate the relevance of effects than to solely rely on significance. By doing so, greater sample sizes will lead to better estimation without introducing over-interpretation of significance.

Finally, the dataset included misses. Due to dropouts resulting from the longitudinal study setting and incomplete datasets, the sample size decreased from $N = 604$ participants at V1 to $n = 330$ participants at V2. Reasons for data exclusion may have correlation to cognitive ability and, thus, imply a selection bias for the remaining $n = 330$ participants, which indicates that the current study may have excluded such participants who had particularly bad courses. As a result, generalisability of results presented in this

methods-focused article may be considerably impaired. Within the remaining data set, only 0.0048% were missing, implying no substantial influence on the estimation procedure within the SEM analysis.

Finally, as only $n = 330$ participants remained in the model, this study relied on the lower bound of necessary data to address such models as discussed in the current manuscript, even though more complex models such as second-order latent growth curve models would be superior as they would be able to model the within-subject nature of the data more properly. Nonetheless, to our understanding, this sample size was sufficient for whole-model comparisons within the current approach.

Conflicts of interest

Jürgen Deckert is the co-recipient of a grant of the Bavarian State Government to BioVariance and an investigator in a European grant to P1Vital.

Author contributions

Sophia Haberstumpf and André Forster equally contributed to data analysis and wrote the initial edition of the manuscript. Jonas Leinweber, Martin Lauer, and Thomas Polak executed and managed medical screenings. Martin Lauer, Thomas Polak, Jürgen Deckert, and Martin J. Herrmann made substantial contributions to the concept and design. Stefanie Rauskolb and Michael Sendtner conducted neurobiological pre-analyses. The draft was critically revised by Martin J. Herrmann, Jonas Leinweber, Stefanie Rauskolb, Johannes Hewig, Michael Sendtner, Martin Lauer, Thomas Polak, and Jürgen Deckert. All authors agreed to the final version.

Acknowledgements

This study was kindly supported by a research grant from the "Vogel Stiftung Dr. Eckernkamp". The authors thank their study nurses Stefanie Karl and Nina Weißenberger for their professional participant recruitment and data collection. We thank Andrea Katzorke, Julia B. M. Zeller, and Laura D. Müller for their perseverance in data collection. We further thank Michaela Kessler, who performed biological analyses. Moreover, the authors thank Inge Gröbner for her dependable support with organizational tasks. Open access funding enabled and organized by ProjektDEAL.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Abbott, A. (2011). Dementia: a problem for our age. *Nature*, *475*, S2–S4.

American Psychiatric Association. (2014). *Diagnostisches und statistisches manual psychischer Störungen–DSM-5®*. Göttingen, Germany: Hogrefe Verlag.

Arnáiz, E., & Almkvist, O. (2003). Neuropsychological features of mild cognitive impairment and preclinical Alzheimer's disease. *Acta Neurologica Scandinavica. Supplementum*, *179*, 34–41.

Avila, J. F., Rentería, M. A., Witkiewitz, K., Verney, S. P., Vonk, J. M. J., & Manly, J. J. (2020). Measurement invariance of neuropsychological measures of cognitive aging across race/ethnicity by sex/gender groups. *Neuropsychology*, *34*(1), 3–14. https://doi.org/10.1037/neu0000584.

Aschenbrenner, S., Tucha, O., & Lange, K. W. (2000). *Regensburger Wortflüssigkeits-Test: RWT*. Göttingen: Hogrefe, Verlag für Psychologie.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). Ime4: Linear Mixed-EffectsMoels Using Eigen and S4. In.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck Depression inventory-II. *San Antonio*, *78*, 490–498.

Berndt, A. E., & Williams, P. C. (2013). Hierarchical regression and structural equation modeling: two useful analyses for life course research. *Family and Community Health*, *36*, 4–18.

Bertola, L., Benseñor, I. M., Gross, A. L., Caramelli, P., Barreto, S. M., Moreno, A. B., . . . Suemoto, C. K. (2021). Longitudinal measurement invariance of neuropsychological tests in a diverse sample from the ELSA-Brasil study. *Brazilian Journal of Psychiatry*, *43*, 254–261.

Bickel, H. (2001). Dementia in advanced age: estimating incidence and health care costs. *Zeitschrift Fur Gerontologie Und Geriatrie*, *34*, 108–115.

Byrne, B. M., & Crombie, G. (2003). Modeling and testing change: an introduction to the latent growth curve model. *Understanding Statistics*, *2*, 177–203.

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

Cooper, C., Sommerlad, A., Lyketsos, C. G., & Livingston, G. (2015). Modifiable predictors of dementia in mild cognitive impairment: a systematic review and meta-analysis. *American Journal of Psychiatry*, *172*, 323–334.

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*, 742–756.

Felt, J. M., Depaoli, S., & Tiemensma, J. (2017). Latent growth curve models for biomarkers of the stress response. *Frontiers in Neuroscience*, *11*, 315.

Fimm, B., & Zimmermann, P. (2001). Testbatterie zur Aufmerksamkeitsprüfung (TAP). In: Version.

Gao, C., Shi, D., & Maydeu-Olivares, A. (2020). Estimating the maximum likelihood root mean square error of approximation (RMSEA) with non-normal data: a Monte-Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 192–201.

Gao, S., Mokhtarian, P. L., & Johnston, R. A. (2008). Nonnormality of data in structural equation models. *Transportation Research Record*, *2082*, 116–124.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in Mplus and SAS. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 676–701.

Haberstumpf, S., Seidel, A., Lauer, M., Polak, T., Deckert, J., & Herrmann, M. J. (2020). Neuronal correlates of the visual-spatial processing measured with functional near-infrared spectroscopy in healthy elderly individuals. *Neuropsychologia*, *148*, 107650.

Härting, C., Markowitsch, H., Neufeld, H., Calabrese, P., Deisinger, K., & Kessler, J. (2000). *WMS-R Wechsler gedächtnistest—revidierte fassung*. Bern, Switzerland: Hans Huber.

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

Hayden, K. M., Kuchibhatla, M., Romero, H. R., Plassman, B. L., Burke, J. R., Browndyke, J. N., & Welsh-Bohmer, K. A. (2014). Pre-clinical cognitive phenotypes for Alzheimer disease: a latent profile approach. *American Journal of Geriatric Psychiatry*, *22*, 1364–1374.

Helmstaedter, C., Lendt, M., & Lux, S. (2001). *Verbaler Lern-und Merkfähigkeitstest: VLMT; Manual*. Göttingen: Beltz-Test.

Hendrix, J. A., Finger, B., Weiner, M. W., Frisoni, G. B., Iwatsubo, T., Rowe, C. C., . . . Carrillo, M. C. (2015). The worldwide alzheimer's disease neuroimaging initiative: an update. *Alzheimer's and Dementia: the Journal of the Alzheimer's Association*, *11*, 850–859.

Jahn, H. (2013). Memory loss in Alzheimer's disease. *Dialogues in Clinical Neuroscience*, *15*, 445–454.

Katzorke, A., Zeller, J. B. M., Müller, L. D., Lauer, M., Polak, T., Deckert, J., & Herrmann, M. J. (2018). Decreased hemodynamic response in inferior frontotemporal regions in elderly with mild cognitive impairment. *Psychiatry Res Neuroimaging*, *274*, 11–18.

Katzorke, A., Zeller, J. B. M., Müller, L. D., Lauer, M., Polak, T., Reif, A., . . . Herrmann, M. J. (2017). Reduced activity in the right inferior frontal gyrus in Elderly APOE-E4 carriers during a verbal fluency task. *Frontiers in Human Neuroscience*, *11*, 46.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*, Vol. *2*. New York, NY: Guilford.

Koscik, R. L., Berman, S. E., Clark, L. R., Mueller, K. D., Okonkwo, O. C., Gleason, C. E., . . . Johnson, S. C. (2016). Intraindividual cognitive variability in middle age predicts cognitive impairment 8–10 years later: results from the wisconsin registry for Alzheimer's prevention. *Journal of the International Neuropsychological Society*, *22*, 1016–1025.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. *R Package Version*, *2*, 734.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Lai, K. (2018). Estimating standardized SEM parameters given nonnormal data and incorrect model: methods and comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 600–620.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*, 1–27.

Ma, Y., Carlsson, C. M., Wahoske, M. L., Blazel, H. M., Chappell, R. J., Johnson, S. C., Asthana, S., & Gleason, C. E. (2021). Latent Factor Structure and Measurement Invariance of the NIH Toolbox Cognition Battery in an Alzheimer's Disease Research Sample. *J Int Neuropsychol Soc*, *27*(5), 412–425. https://doi.org/10.1017/s1355617720000922

Makkar, S. R., Lipnicki, D. M., Crawford, J. D., Kochan, N. A., Castro-Costa, E., Lima-Costa, M. F., . . . Sachdev, P. (2020). APOE ε4 and the influence of sex, age, vascular risk factors, and ethnicity on cognitive decline. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *75*, 1863–1873.

Mayeux, R. (2010). Early Alzheimer's disease. *New England Journal of Medicine*, *362*, 2194–2201.

Meyers, J. E., & Meyers, K. R. (1996). *Rey Complex Figure Test and Recognition Trial Supplemental Norms for Children and Adults*. Psychological Assessment Resources.

Moreira, P. S., Santos, N., Castanho, T., Amorim, L., Portugal-Nunes, C., Sousa, N., & Costa, P. (2018). Longitudinal measurement invariance of memory performance and executive functioning in healthy aging. *PLoS One*, *13*, e0204012.

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. *J Int Neuropsychol Soc*, *18*(6), 1071–1080. https://doi.org/10.1017/s1355617712000859

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

National Institute of Mental Health. (2011). NIMH Research Domain Criteria (RDoC).

Nestor, P. J., Fryer, T. D., & Hodges, J. R. (2006). Declarative memory impairments in Alzheimer's disease and semantic dementia. *NeuroImage*, *30*, 1010–1020.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*, 587–598.

Park, D. C., & Festini, S. B. (2017). Theories of memory and aging: a look at the past and a glimpse of the future. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *72*, 82–90.

Petersen, R. C. (2000). Mild cognitive impairment: transition between aging and Alzheimer's disease. *Neurologia*, *15*, 93–101.

Polak, T., Herrmann, M. J., Müller, L. D., Zeller, J. B. M., Katzorke, A., Fischer, M., … Deckert, J. (2017). Near-infrared spectroscopy (NIRS) and vagus somatosensory evoked potentials (VSEP) in the early diagnosis of Alzheimer's disease: rationale, design, methods, and first baseline data of the Vogel study. *Journal of Neural Transmission*, *124*, 1473–1488.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's and Dementia: the Journal of the Alzheimer's Association*, *9*, 63–75.e62.

R Core Team. (2016). Vienna: R Foundation for Statistical Computing.

Rahmadi, R., Groot, P., van Rijn, M. H., van den Brand, J. A., Heins, M., Knoop, H., & Heskes, T. (2018). Causality on longitudinal data: stable specification search in constrained structural equation modeling. *Statistical Methods in Medical Research*, *27*, 3814–3834.

Riedel, W. J., & Blokland, A. (2015). Declarative memory. *Handbook of Experimental Pharmacology*, *228*, 215–236.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*, 1–36.

Rowe, J. B. (2010). Connectivity analysis is essential to understand neurological disorders. *Frontiers in Systems Neuroscience*, *4*, 144.

Sayegh, P., & Knight, B. G. (2014). Functional assessment and neuropsychiatric inventory questionnaires: measurement invariance across hispanics and non-Hispanic whites. *Gerontologist*, *54*(3), 375–386. https://doi.org/10.1093/geront/gnt026

Schmitt, N., Golubovich, J., & Leong, F. T. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: an illustrative example using Big Five and RIASEC measures. *Assessment*, *18*, 412–427.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Statistisches Bundesamt (2018). *Bildungsstand der Bevölkerung - Ergebnisse des Mikrozensus 2017*. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsstand/Publikationen/Downloads-Bildungsstand/bildungsstand-bevoelkerung-5210002177004.pdf;jsessionid=356E153EF23D5F8006CEC57A4B85AF16.internet8731?__blob=publicationFile

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. Northridge, CA: Harper Collins.

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

Tuokko, H. A., Chou, P. H., Bowden, S. C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *J Int Neuropsychol Soc*, *15*(3), 416–425. https://doi.org/10.1017/s1355617709090602

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, *30*, 1006–1016.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363.

Williams, B. D., Chandola, T., & Pendleton, N. (2018). An application of Bayesian measurement invariance to modelling cognition over time in the English longitudinal study of ageing. *International Journal of Methods in Psychiatric Research*, *27*, e1749.

Winblad, B., Wimo, A., Engedal, K., Soininen, H., Verhey, F., Waldemar, G., … Schindler, R. (2006). 3-year study of donepezil therapy in Alzheimer's disease: effects of early and continuous therapy. *Dementia and Geriatric Cognitive Disorders*, *21*, 353–363.

World Health Organization [WHO] (2016). *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. Geneva, Switzerland: World Health Organization.

World Health Organization [WHO] (2019). International Classification of Diseases (ICD). Version updated in 2019. World Health Organization. Retrieved from https://icd.who.int/browse10/2019/en

World Medical Association (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*, *310*, 2191–2194.

Xu, W., Tan, L., Wang, H. F., Jiang, T., Tan, M. S., Tan, L., ... Yu, J. T. (2015). Meta-analysis of modifiable risk factors for Alzheimer's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, *86*, 1299–1306.

Yilmaz, F. N. (2019). Comparison of different estimation methods used in confirmatory factor analyses in non-normal data: a Monte Carlo study. *International Online Journal of Educational Sciences*, *11*, 131–140.

Zeller, J. B. M., Katzorke, A., Müller, L. D., Breunig, J., Haeussinger, F. B., Deckert, J., ... Herrmann, M. J. (2019). Reduced spontaneous low frequency oscillations as measured with functional near-infrared spectroscopy in mild cognitive impairment. *Brain Imaging and Behavior*, *13*, 283–292.

Appendix A:

Table A1. Mixed model comparisons of declarative memory

Model	AIC	BIC	χ^2	df	p ($>\chi^2$)
DM1	2629.5	2647.1			
DM2	2600.4	2635.5	37.1459	4	1.681E-07***
DM3	2600.9	2644.9	3.4385	2	.1792
DM4	2600.5	2653.2	4.4083	2	.1103
DM5	2604.4	2665.9	0.1603	2	.923

DM = declarative memory; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table A2. Mixed-model comparisons of attention

Model	AIC	BIC	χ^2	df	p ($>\chi^2$)
AT1	1318.9	1336.5			
AT2	1322.9	1358.1	3.9687	4	.41026
AT3	1319.1	1363	7.8371	2	.01987*
AT4	1320.7	1373.4	2.3888	2	.30289
AT5	1322.5	1384	2.2136	2	.33062

AT = attention; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table A3. Mixed-model comparisons of working memory

Model	AIC	BIC	χ^2	df	p ($>\chi^2$)
VSP 1	1859.3	1876.8			
VSP 2	1851.2	1886.4	16.0394	4	.002967**
VSP 3	1850.8	1894.7	4.4825	2	.106325
VSP 4	1848.6	1901.3	6.1641	2	.045866*
VSP 5	1850.8	1912.3	1.762	2	.414375

WM = working memory; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table A4. Mixed-model comparisons of visual–spatial processing

Model	AIC	BIC	χ^2	df	p ($>\chi^2$)
WM1	1222.1	1239.7			
WM 2	1218.2	1253.4	11.9321	4	.01786*
WM 3	1220.3	1264.3	1.862	2	.39415
WM 4	1220.1	1272.8	4.2453	2	.11972
WM 5	1222.8	1284.3	1.3083	2	.51989

VSP = visual–spatial processing; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table A5. Best Mixed-models for the fixed effects of Declarative Memory for the latent factors and composite approach

DM	Estimate	SE	df	t	p	Beta
(Intercept)	−0.165/46.011	0.104/0.982	296/296	−1.582/46.856	.115/<.001***	0.03/0.02
Gender	0.999/4.896	0.21/1.972	296/296	4.765/2.483	<.001***/.014*	0.44/0.22
Time	−0.3/−1.249	0.103/1.194	296/296	−2.898/−1.046	.004**/.296	−0.13/−0.06
Age	−0.16/−0.091	0.068/0.642	296/296	−2.34/−0.142	.020*/.887	−0.11/−0.006
Gender: time	0.518/−1.109	0.208/2.397	296/296	2.49/−0.463	.013*/.644	0.23/−0.05
Time:age	−0.071/0.689	0.068/0.78	296/296	−1.054/0.883	.293/.378	−0.05/0.05

DM = declarative memory; SE = standard error.
Values before the slash refer to the latent factor approach and values after the slash to the composite approach.

Table A6. Best mixed-models for the fixed effects of Attention for the latent factors and composite approach

AT	Estimate	SE	df	*t*	*p*	Beta
(Intercept)	0.052/46.373	0.032/0.314	295/295	1.621/147.806	.106/<.001***	0.01/0.009
Gender	0.128/0.72	0.065/0.637	295/295	1.966/1.131	.050/.259	0.18/0.11
Time	0.034/0.46	0.039/0.27	295/295	0.852/1.707	.395/.089	0.05/0.07
Age	0.008/0.266	0.021/0.205	295/295	0.36/1.296	.719/.196	0.02/0.06
BDNF	−0.002/−0.003	0.001/0.009	295/295	−2.278/−0.288	.023*/.774	−0.1/−0.01
Gender:time	0.017/0.517	0.08/0.548	295/295	0.214/0.944	.831/.346	0.02/0.08
Time:age	0.026/0.252	0.026/0.176	295/295	1.026/1.43	.306/.154	0.06/0.06
Time:BDNF	0.002/−0.001	0.001/0.008	295/295	1.612/−0.091	.108/.927	0.09/−0.004

AT = attention; SE = standard error; BDNF = Brain-Derived Neurotrophic Factor; BDI-II = Beck Depression Inventory-II (Beck et al., 1996).
Values before the slash refer to the latent factor approach and values after the slash to the composite approach.

Table A7. Best mixed-models for the fixed effects of Visual–Spatial Processing for the latent factors and composite approach

VSP	Estimate	SE	df	*t*	*p*	Beta
(Intercept)	−0.383/57.283	0.062/0.931	294/294	−6.195/61.531	<.001***/ <.001***	−0.02/−0.02
Gender	−0.272/−5.151	0.127/1.905	294/294	−2.15/−2.704	.032*/.007**	−0.21/−0.24
Time	−0.527/−1.325	0.049/1.116	294/294	−10.662/−1.188	<.001***/.236	−0.41/−0.06
Age	0.002/1.356	0.04/0.609	294/294	0.052/2.227	.958/.027	0.003/0.1
BDNF	0.002/0.028	0.002/0.027	294/294	1.192/1.06	.234/.290	0.06/0.05
BDI-II	−0.022/−0.155	0.011/0.162	294/294	−2.003/−0.955	.046*/.340	−0.1/−0.04
Gender:time	−0.302/−3.384	0.101/2.283	294/294	−2.987/−1.482	.003**/.139	−0.24/−0.16
Time:age	0.041/1.715	0.032/0.73	294/294	1.283/2.35	0.20*/.019*	0.05/0.12
Time:BDNF	0.002/0.012	0.001/0.032	294/294	1.469/0.369	.143/.712	0.06/0.02
Time:BDI-II	−0.012/−0.048	0.009/0.194	294/294	−1.443/−0.249	.150/.803	−0.06/−0.01

VSP = visual–spatial processing; SE = standard error; BDNF = Brain-Derived N eurotrophic Factor; BDI-II = Beck Depression Inventory-II (Beck et al., 1996).
Values before the slash refer to the latent factor approach, values after the slash to the composite approach.

Table A8. Best mixed-models for the fixed effects of Working Memory for the latent factors and composite approach

WM	Estimate	SE	df	*t*	*p*	Beta
(Intercept)	−0.012/64.297	0.046/1.148	296/296	−0.266/56.019	.791/<.001***	0.02/0.009
Gender	0.233/2.518	0.093/2.305	296/296	2.51/1.092	.013*/.276	0.28/0.11
Time	−0.021/4.517	0.023/0.99	296/296	−0.905/4.562	.366/<.001***	−0.02/0.2
Age	−0.046/−0.377	0.03/0.75	296/296	−1.507/−0.503	.133/.616	−0.08/−0.02
Gender:time	−0.074/−1.208	0.046/1.989	296/296	−1.597/−0.608	.111/.544	−0.09/−0.05
Time:age	0.002/0.361	0.015/0.647	296/296	0.13/0.558	.896/.577	0.004/0.02

WM = working memory; SE = standard error.
Values before the slash refer to the latent factor approach, values after the slash to the composite approach.