Published in final edited form as:
Psychol Assess. 2016 December ; 28(12): 1576–1585. doi:10.1037/pas0000285.

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)

N. Maritza Dowling,
University of Wisconsin-Madison, Wisconsin Alzheimer's Disease Research Center

Daniel M. Bolt, and
University of Wisconsin-Madison

Sien Deng
University of Wisconsin-Madison, Wisconsin Alzheimer's Disease Research Center

Abstract

When assessments are primarily used to measure change over time, it is important to evaluate items according to their sensitivity to change, specifically. Items that demonstrate good sensitivity to between-person differences at baseline may not show good sensitivity to change over time, and vice versa. In this study, we applied a longitudinal factor model of change (LFMC) to a widely-used cognitive test designed to assess global cognitive status in dementia, and contrasted the relative sensitivity of items to change. Statistically-nested models were estimated introducing distinct latent factors related to initial status differences between test-takers and within-person latent change across successive time points of measurement. Models were estimated using all available longitudinal item-level data from the Alzheimer's Disease Assessment Scale-Cognitive section (ADAS-Cog), including participants representing the full-spectrum of disease status who were enrolled in the multi-site Alzheimer's Disease Neuroimaging Initiative (ADNI). Five of the thirteen ADAS-Cog items demonstrated noticeably higher loadings with respect to sensitivity to change. Attending to performance change on only these five items yielded a clearer picture of cognitive decline more consistent with theoretical expectations in comparison to the full thirteen-item scale. Items that show good psychometric properties in cross-sectional studies are not necessarily the best items at measuring change over time, such as cognitive decline. Applications of the methodological approach described and illustrated in this study can advance our understanding regarding the types of items that best detect fine-grained early pathological changes in cognition.

Corresponding author: N. Maritza Dowling, Ph.D., Department of Biostatistics and Medical Informatics, University of Wisconsin, School of Medicine and Public Health, Madison, WI. 53792, USA; Phone: 608-263-4100; nmdowlin@biostat.wisc.edu.

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

Disclosures: There were no actual or potential conflicts of interest for any of the authors.

Keywords

Longitudinal Factor Model; Item Sensitivity to Change; Measurement Invariance; ADAS-Cog; Cognitive Decline; Structural Equation Modeling; Alzheimer's Disease

Introduction

Neurocognitive testing is a vital component to early detection, accurate diagnosis, and the monitoring of disease progression and evaluation of treatment effects. The measurement of cognitive abilities through testing is also critical in assessing the effectiveness and clinical meaningfulness of symptomatic and disease-modifying clinical trials. Accurate clinical diagnosis and classification of individuals into the correct disease status depends, to some extent, on the validity of the cognitive test score interpretations. The validity of cognitive test scores can be measured by their capacity to detect patterns of cognitive deficits that might be indicative of abnormal decline (Bondi et al., 1994; Bondi et al., 1995; Salmon & Lange, 2001). It is important to distinguish between pathological cognitive decline or “impairment” and normal (non-pathological) age-associated cognitive decline. The first can be measured by performance on cognitive tests well below normative standards while the latter refers to a change from baseline or expected ability compared to established age-group norms for the test (Slick, 2006). In either case, it is fairly standard to use 1 to 1.5 SD change from estimated baseline or previous performance as the marker of a meaningful “change”.

Outcome measures that detect subtle and fine changes over time in cognitive function are undoubtedly useful to enhance our understanding of decline in cognition due to disease progression and to accurately assess the clinical benefits of therapeutic interventions. Optimally-developed cognitive measures that are sensitive to the early emergence of clinical symptoms may also prove to be highly associated with underlying brain pathology providing strong evidence of biomarker validation (Sperling et al., 2011). Consequently, neuropsychological outcome measures that capture change and reliably detect “true” residual cognitive function as a clinical indicator of remaining intact cerebral architecture (or brain structure) are as important as understanding the clinical effects of underlying pathology (Bussire et al., 2003; Esiri & Chance, 2012). Several recent publications have demonstrated that improved cognitive composite scores using multivariate machine learning approaches to weight individual test items (Llano, Laforet, & Devanarayan, 2011) or combining items from different scales (Skinner et al., 2012) can perform as well as or better than putative neuroimaging or cerebrospinal fluid (CSF) biomarkers in predicting conversion to dementia.

The utility of a cognitive test primarily used to measure change over time can be evaluated in terms of the differential contribution and sensitivity of individual items to change. Items in cognitive assessment instruments that demonstrate good psychometric properties and sensitivity to between-person baseline or cross-sectional differences may not show as much sensitivity to change over time. The converse may also be true. Concerns about accurate test scoring, appropriate methodology for analyzing test score data, and optimal selection of items to assess cognitive and functional abilities for disease prognosis may explain the increasingly popular application of advanced latent variable modeling or item response theory (IRT) methodology in health sciences and aging research (e.g., Crane et al., 2008; Flynn, Dombeck, DeWitt, Schulman, & Weinfurt, 2008; Fries, Bruce, & Cella, 2005; Mungas, & Reed, 2000; Mungas, Reed, & Kramer, 2003; Proust-Lima, Amieva, Dartigues, & Jacqmin-Gadda, 2007; Salsman et al., 2014). In IRT modeling, item fitness is generally evaluated on an item-by-item basis using the item responses across all test takers. The selection of candidate items for test assembly is generally guided by item properties such as item difficulty or sensitivity to discriminate between individuals with different levels of the underlying trait measured by the test. When such tests are ultimately used to study change over time, it is often implicitly assumed that the change that occurs is best understood as change in the same latent trait that distinguished individuals at baseline. However, the introduction of different types of latent variables in latent variable models makes it possible to test this assumption. If violated, the model can also provide a basis for separate evaluations of individual items according to their relative sensitivities to between- and within-person (change) factors, which can be valuable in identifying items that increase the usefulness and responsiveness of neuropsychological and neurobehavioral assessments designed to measure change. Such analyses provide both theoretical and practical value in the design and analysis of scale instruments that may seek to study both aspects of cognitive performance in an individual.

Meredith and Horn (2001) proposed a longitudinal factor model of change (LFMC) containing specific latent factors related to baseline cognitive status differences between test takers and latent change across the measurement time points. A key advantage of structural factor models, such as the LFMC, is the ability to sequentially impose constraints on the parameters of interest allowing the study of the invariance of specific covariances and/or factor loadings of items in a test over time (McArdle, 2007). Standard measurement equivalence or measurement invariance analyses, as described in Meredith (1993), can be applied to the model allowing the detection of individual test items that display significantly greater (or lesser) sensitivity to change as compared to their sensitivity to initial differences.

In theory, the LFMC model thus also provides a mechanism by which the lack of measurement invariance for a given assessment scale over time can be explained. Traditional forms of measurement invariance analysis start with an attempt to confirm that a scale measures the same latent trait over time. As this assumption is frequently violated, it becomes important to not only understand why, but to introduce methodological alternatives that allow such changes to occur. By introducing separate between- versus-within person latent factors, the LFMC model provides such a mechanism. The purpose of this study was to illustrate the application of the LFMC modeling framework to identify items most sensitive to change across time using data from the Alzheimer’s Disease Assessment Scale-Cognitive section (ADAS-Cog) (Rosen, Mohs, & Davis, 1984). The ADAS-Cog is a rating instrument commonly used to measure cognitive dysfunction in clinical trials and for detecting, tracking, and staging AD (e.g., Aisen et al., 2008; Feldman et al., 2010; Kurz, Farlow, Quarg, & Spiegel, 2004; Rafii et al., 2011; Sano et al., 2011; Shah et al., 2013; Suzuki et al., 2013). ADAS-Cog scores are obtained from written and verbal responses to items measuring key cognitive domains typically affected in AD including verbal episodic memory, language, comprehension, and ideomotor praxis.

Method

Participants

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI has been to test whether imaging and other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness as well as lessen the time and cost of clinical trials. To date ADNI protocols have recruited over 1,400 adults, ages 55 to 90, clinically diagnosed as cognitively normal (CN), MCI, or AD. For up-to-date information on ADNI protocols, see www.adni-info.org.

The population for the present study consisted of all participants with available item-level data on the ADAS-Cog up to the 24-month follow-up visit. A period of two years at one year intervals was considered a reasonable amount of time to study cognitive decline in this sample (see, e.g., De Jager, Blackwell, Budge, & Sahakian, 2005; Marquis et al. 2002). The sample included individuals across the full spectrum of disease status, from CN, early and late MCI, and early AD. Details on the study protocol for the clinical diagnosis of participants into disease categories can be found at http://www.adni-info.org/Scientists/ADNIStudyProcedures.aspx. Briefly, clinical diagnosis is established in a multi-step process combining results from neuropsychological, neuropsychiatric, and functional tests and clinical judgement based on published criteria for dementia. As shown in Table 1, the analytical sample for fitting the LFCM model included 1,217 older adults diagnosed at study entry as CN (N = 343), MCI (N = 764), and AD (N = 110). This represented 82% (1,217/1,480) of the total sample recruited in the ADNI studies at the time the data for the present study were downloaded. At month 12, we retained 90% of the analytical sample (1,094/1,217), and at month 24, roughly 60% (733/1,217) had item-level data on the ADAS-Cog-13. Reasons for lost to follow-up or missing appointments in ADNI vary by biomarker. The most common reasons reported in the literature are death, cognitive impairment, depression, and other health-related complications (Lo et al., 2012). Participants in the current study were mostly male (57%), ranged in age from 55 to 91 years (M = 73, SD = 7.08), and reported an average of 16.11 years of education (SD = 2.75; range, 6–20 years). Table 1 also reports global cognitive function at baseline measured by the Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) and the Clinical Dementia Rating Scale Sum of Boxes (CDR-SB; Morris, 1993).

Instrument

The classic or standard ADAS-Cog (Rosen, Mohs, & Davis 1984) includes 11 items and the revised and expanded ADAS-Cog-13 (Mohs et al., 1997) includes two additional items measuring visual attention and concentration (digit or number cancellation) and delayed verbal recall. Total test scores may range from 0 to 70 in the standard version of the test and from 0 to 85 in the expanded version with lower scores indicating better cognitive performance. Items were scored following the test developers guidelines. We studied item sensitivity to change using the expanded version of the scale denoted here as ADAS-Cog-13. Table 2 lists the 13 items by cognitive domain and scoring scheme.

The ADNI study administers alternate test forms at each visit in which only the word lists are varied to minimize practice effects. To insure unambiguous interpretation of changes in the ADAS-Cog-13 between the studied time points, we conducted longitudinal measurement invariance tests over a 24-month interval to determine whether the test items assessed the same attribute across time (Horn, & McArdle, 1992; Meredith, 1993). Longitudinal invariance was evaluated using a multi-group confirmatory factor analysis within the framework of structural equation modeling (SEM; Meredith, 1993; Schaie, Maitland, Willis, & Intrieri, 1998). We assessed the degree to which ADAS-Cog-13 factor structure (configural invariance), factor loadings (metric invariance), factor variance/covariance and item means (scalar invariance), and item error variances were similar across time. The results provided evidence in support of the test's longitudinal factorial invariance over the 24-month period. (Results are available upon request from the first author.) Means and standard deviations of the ADAS-Cog-13 at baseline are also reported in Table 1.

The Longitudinal Factor Model of Change

Using item-level cognitive outcome scores, the LFMC model was applied incorporating specific latent factors to measure cross-sectional differences in cognitive functioning (between participants at baseline) and the sensitivity to capture within-person change in cognitive functioning over time. Assuming that cognitive functions are repeatedly measured across $t = 3$ time points, the general longitudinal factor model of change is specified as:

$$\begin{gathered} x_{ij1}=\lambda_{j1}f_{i1}+\varepsilon_{ij1} \\ x_{ij2}=\lambda_{j1}f_{i1}+\lambda_{j2}f_{i2}+\varepsilon_{ij2} \\ x_{ij3}=\lambda_{j1}f_{i1}+\lambda_{j2}f_{i2}+\lambda_{j3}f_{i3}+\varepsilon_{ij3} \end{gathered}$$

where x_{ij1}, x_{ij2}, and x_{ij3} denote the item score for person i on item j across t time points, denoted here as 1, 2 and 3, respectively; f_{i1}, f_{i2}, and f_{i3} are latent factors denoting the baseline cognitive status (baseline factor 1), cognitive change from $t = 1$ to $t = 2$ (change factor 2), and cognitive change from $t = 2$ to $t = 3$ (change factor 3) for person i, respectively. In this model specification the baseline factor 1, f_1, represents a factor reflecting differences between persons in baseline cognitive functioning, while change factor 2, f_2, and change factor 3, f_3, represent how an individual has changed across the two respective one-year time intervals. The parameters $\lambda_{j\,1}$, $\lambda_{j\,2}$, *and* $\lambda_{j\,3}$ thus indicate the initial status and change status loadings (with respect to two time intervals) for item j, respectively. Finally, e_{ij1}, e_{ij2}, and e_{ij3} are the respective error terms reflecting the uniqueness variance for each item j at a given time point t.

In this study, we evaluated the general unrestricted model, denoted as the Varying Factor Model, which allows all factor loadings to differ across time ($\lambda_{j1}\neq\lambda_{j2}\neq\lambda_{j3}$), against restricted models in which $\lambda_{j1} = \lambda_{j2} = \lambda_{j3}$ for all items (denoted as the Constant Factor Model), and where only $\lambda_{j2} = \lambda_{j3}$ for all items (specified as the Baseline and Change Factor Model). It is important to note that the Constant Factor Model is the most restrictive of the three models and is statistically equivalent to a model that assumes measurement invariance in a single latent trait over time. The Varying Factor Model is the least restrictive model allowing not only differences in item loadings across the baseline and change factors but also between the two factors representing change. The Baseline and Change Factor Model allows the item loadings to differ between the baseline and change factors but assumes the change factors are invariant. Additionally, all the models included covariance terms between factors, which allows us to see (and account for) the potential relationship between cognitive functioning at baseline and subsequent declines.

The three models were compared to gather statistical evidence regarding the potential differential sensitivities of the items across latent factors. In the application of the LFMC model to the ADAS-Cog item scores illustrated here, we viewed baseline factor 1 (f_{i1}) as representing the level of cognitive functioning observed at time 1. In this case, an item's loading (λ_{j1}) indicates how well the item measures the participants' cognitive status at $t = 1$, while λ_{j2} and λ_{j3} indicate how well it measures change from $t=1$ to $t = 2$ and from $t=2$ to $t = 3$, respectively. Comparing models that differ in the nature of constraints imposed on their factor loadings within the LFMC is conceptually akin to confirmatory factor analysis approaches to establishing measurement invariance (Meredith, 1993; Millsap & Meredith, 2007; Reise, Widaman, & Pugh, 1993). Since the three models of interest possess a fully-nested structure in terms of the nature of the constraints imposed, model comparison can be conducted via either chi-square difference testing or through use of standard model comparison and goodness-of-fit criteria applied in confirmatory factor analysis. The latter criteria included the Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), sample-size-adjusted BIC (SABIC; Sclove, 1987), comparative fit index (CFI; Bentler, & Bonett, 1980), Tucker-Lewis index (TLI; Tucker, & Lewis, 1973), and root mean square error of approximation (RMSEA; Browne & Cudeck, 1992; Steiger, & Lind, 1980). Model comparison provides the basis for claims that item performance may or may not vary in relation to between-person versus within-person factors.

Additional Analyses

To further validate the best model obtained in the previous step and to illustrate the value of distinguishing among the 13 ADAS-Cog items in their capacity to capture cognitive decline, we compare the observed mean item scores at each time point using a subset of the ADAS-Cog items found to be maximally sensitive to change over time. We also compared the average growth estimates (i.e., rate of change) obtained from fitting separate linear latent growth curve models (LGCM) (Bollen & Curran, 2006) using the total score of the full test versus those items found most sensitive to change. To facilitate the comparison of the sensitivities of different sets of items in capturing change, these analyses were conducted using cases with complete values in all items across all the observation time points evaluated. This produced an analytical sample of 491 cases with non-missing data. All models were estimated using *MPlus 7.31* (Muthén, & Muthén, 1998–2015). All descriptive statistics and graphics were obtained with the programming language R, Version 3.2.0 (R Core Team, 2015).

Results

Item sensitivity using the Longitudinal Factor Model of Change

Chi-square difference (χ^2_{diff}) tests (also known as likelihood ratio tests) were used to compare the fit of the competing nested models. The hypothesis test comparing the Baseline and Change Factor Model vs the Varying Factor Model was not rejected, $\chi^2_{diff}\,(12, N{=}1,217){=}8.16$, $p = .77$, suggesting that the more parsimonious model (Baseline and Change Factor Model) was preferred. Conversely, the test comparing the Baseline and Change Factor Model with the Constant Factor Model rejected the null hypothesis $\chi^2_{diff}\,(13, N{=}1,217){=}3.48$, $p < .001$, implying that the model containing more freely estimated parameters (in this case the Baseline and Change Factor Model) provided a better fit than the Constant Factor Model. A summary of the fit produced by each sequential LFMC model using the specified criteria is shown in Table 3. With the exception of the AIC, which is known to favor more complex models (West, Taylor, and Wu, 2014), the Baseline and Change Factor Model outperformed the Constant and the Varying Factor models. Therefore, the Baseline and Change Factor Model was selected as the most appropriate model in our study. As change factors 2 and 3, respectively, represent the sensitivity to capturing an individual's cognitive change from time 1 to time 2 and time 2 to time 3, the results suggest that this sensitivity is relatively stable over time (i.e., $\lambda_{j2}{=}\lambda_{j3}$ in the Baseline and Change Factor Model), but different from the relative sensitivity to baseline differences in cognitive functioning.

Tables 4 and 5 display the standardized estimates of item loadings obtained from the baseline factor 1 and growth or latent change factors 2 and 3 for all items in the ADAS-Cog-13. In the left columns of both tables, we can see that item 1 (*word-recall-trial 1 to 3*), item 4 (*delayed word recall*), and item 8 (*word-recognition*) show the largest loadings with respect to Factor 1, which suggests that they are optimal in discriminating participants' cognitive levels in their initial cognitive status. In contrast, items 2 (*commands*), 9 (*remember instructions*), and 12 (*spoken language*) were not as discriminating. However, items 2 (*commands*), 9 (*remember instructions*), 10 (*comprehension*), 11 (*word finding*), and 12 (*spoken language*) had the largest loadings on the change factors 2 and 3. This indicates their relatively stronger sensitivity to an individual's change or decline in cognitive ability across the three time points. With the exception of item 2 (*commands*), all the items identified as most sensitive were clinician-rated items measuring language (items, 2, 10, 11, and 12) and memory (item 9) abilities. Interestingly, items 1 (*word-recall-trial 1 to 3*), 3 (*copy geometric forms*), 4 (*delayed word recall*), and 8 (*word-recognition*), designed to assess predominantly memory skills, had the lowest loadings on change factors 2 and 3, suggesting that among the 13 items on the test, these items showed the least sensitivity to changes across time.

Figure 1 illustrates the relationship between item factor loadings and differential functioning across the two types of estimated latent factors. For example, items 1 (*word-recall*), 4 (*delayed word recall*), and 8 (*word-recognition)* perform best at distinguishing the cognitive functioning between individuals at their baseline status, but show the least sensitivity to capture the individual's cognitive decline over time. On the other hand, items 9 (*remember* *instructions*), 10 (*comprehension*), and 12 (*spoken language*) show the greatest sensitivity to within-individual change in cognitive ability longitudinally, but they are not as sensitive as the other items in measuring the initial cognitive level. If we apply, for example, 0.35 as a minimum factor loading "rule of thumb" for item selection (cf; Stevens, 2009), we can identify some items that perform relatively well both in sensitivity to baseline differences and change across time. That is, if an item has a factor loading greater than 0.35 on both latent factors, then it can be viewed as a good item in terms of capturing both initial differences and change over time. Examples of such items in the ADAS-Cog 13 are items 5 (*naming objects*) and 7 (*orientation*). In general terms, these findings suggest that items that appear to function well in distinguishing baseline between-person differences in the underlying trait measured by the instrument might not necessarily function as well for detecting within-person differences. In fact, we found a strong negative correlation between the baseline factor 1 and the change factor 2, $r(11) = -0.85$, $p < .001$.

Comparative Analyses

Table 6 shows the observed mean item scores at each time interval for the following three groups of ADAS-Cog-13 items: 1) a five-item group containing the items demonstrating greatest sensitivity to change, 2) the remaining eight-item group of non-selected items, and 3) the full thirteen-item group. As indicated above, the five-item group was comprised of the items with the largest loadings on change factor 2, and hence had the greatest sensitivity to detect cognitive decline. These included *commands*, *remember instructions*, *comprehension*, *word finding*, and s*poken language*. Figure 2 depicts the trends of the observed mean scores over the three data collection time points with a 95% confidence interval. We can see that only the trajectory of the five-item total score outcome displayed a clear increasing trend (greater cognitive impairment) across all three time points, consistent with a linear cognitive decline progression. Conversely, mean scores across time based on the thirteen- and eight-item outcome measures show an inconsistent pattern, initially showing improvement from $t = 1$ to $t = 2$, and then a decline from $t = 2$ to $t = 3$. This finding may simply be due to noise introduced by items that are not sensitive to within-person change. However, it is also possible that such results reflect some form of practice effects on particular items; effects that potentially become offset by cognitive declines at later time points. This interpretation seems particularly plausible given that the pattern of improvement is most noticeable for the CN population, less so for the mild MCI group, and not at all for the AD group, where the effects of actual cognitive decline have likely offset any benefit from practice.

A summary of the estimated average linear growth parameters (i.e., mean slope or rate of change) obtained from an unconditional LGCM applied separately to the five-, eight-, and thirteen-item longitudinal outcome measures is presented in Table 7. To facilitate comparison across item groups**,** the table shows the standardized average growth slope estimates. The results again support the five-item aggregate score as the most sensitive measure of change. The estimated growth parameter for the five-item outcome measure was the highest, also yielding the smallest standard error and most significant *p*-value.

Discussion

The purpose of this study was to describe and illustrate the application of the LFMC framework to serial item-level data from the ADAS-Cog-13 to examine the relative sensitivities of items to detect cross-sectional differences in cognitive function between participants and within-person change in cognition over time. The model identified five items indicating a progressive decline across the three time points. To validate the LFMC results we used longitudinal aggregate scores as outcomes to compare the performance of the five model-selected items in capturing change to that of non-selected items and all thirteen items in the test. We also used a LGC modeling approach to estimate and compare growth factors across item-groups. These analyses supported the superiority of longitudinal aggregate scores obtained from the five items showing greater sensitivity to change, as the entire collection of 13 items failed to display a progressive decline. Most factor analytic studies of longitudinal measures attend solely to between-person factors when interpreting what the items are measuring. One limitation of this class of analyses is that the frequently observed lack of measurement invariance with respect to such factors is difficult to explain. An appealing feature of the model studied in this paper is that it provides an explanation for the lack of measurement invariance over time, namely that items are disproportionately sensitive to cognitive decline.

As evidenced by numerous publications, there is an urgent need to develop cognitive outcome measures that are sensitive in revealing early synaptic and neuronal dysfunction associated with AD pathology (Becker, Greig, & Giacobini, 2008; Broich, Schlosser-Weber, Weiergraber, & Hampel, 2012; Robert et al., 2010; Sperling et al., 2011). Optimized measures of different cognitive processes are a necessary component of clinical trials of drug-modifying therapies for accurate assessment of treatment benefits. More sensitive primary or co-primary cognitive outcomes may also lower the cost and enhance the efficiency of clinical trials by reducing sample size requirements and increasing power to detect significant differences at both the individual and aggregate sample levels. The evaluation protocol of items for cognitive tests should involve not only their difficulty level and ability to discriminate amongst individuals across the full spectrum of the disease, but also the extent to which they are sensitive to detect change over time; especially when the objective of the assessment is to measure change and gain better understanding of intra-individual processes. A latent variable modeling framework which links longitudinal item-level observations to latent variables or factors allowing the distinction between item sensitivity to baseline interindividual differences and intraindividual change provides one such tool. We demonstrated that desirable cross-sectional psychometric properties of items that made up a test do not necessarily translate into desirable item characteristics for the analysis of change.

To some extent, the overall quality of a test to measure change over time stems from the quality of its items to detect such change. In close agreement with previous cross-sectional studies applying IRT models to the analysis of item-level data from ADAS-Cog (e.g., Benge, Balsis, Geraci, Massman, & Doody, 2009; Ueckert, et al., 2014), we found that items measuring episodic memory deficits (e.g., word recall, delayed word recall, and word recognition) performed very well in differentiating respondents with different levels of the cognitive ability measured by the test. Yet, the performance of the same items relative to other items in the test was not as optimal in measuring individual differences in change over time. We found that items assessing predominantly language or semantic memory (e.g., spoken language, comprehension of spoken language, commands, word finding, and naming objects) were much more sensitive or responsive to capturing intraindividual differences in change over time than those measuring episodic memory. These results suggest that the inclusion of *selected* items that detect changes in semantic memory may enhance the usefulness of a test to identify and track true rate of cognitive decline over time. In fact, previous cross-sectional studies (e.g., Welsh, Butters, Hughes, Mohs, & Heyman, 1991; 1992) have indicated that neuropsychological tests measuring learning and memory are highly sensitive for detecting very mild cases of AD, but proved of little value for detecting changes in the disease process. In contrast, the authors found that the performance on lexical-semantic processing measures was a better indicator of changes in clinical diagnosis and disease status. Other studies utilizing longitudinal designs and large community-based non-demented samples at study entry (Amieva, et al., 2008; Wilson, Leurgans, Boyle, & Bennett, 2011) reported that the first measurable cognitive decline was obtained on cognitive tests assessing semantic memory and conceptual formation becoming evident as early as 12 years before conversion to dementia. Although these studies provide support to our findings on the role of semantic memory measures on discriminating disease progression, the focus of our analysis was chiefly on identifying specific items in a scale showing the highest sensitivity to possibly pathological cognitive changes in the underlying trait measured by the scale regardless of the person clinical diagnosis.

More applications and extensions of the LFMC modeling framework exist that may enhance the value of this methodology as an effective tool to monitor and assess cognitive changes. In this study, we have only considered changes using three waves of data to facilitate the introduction and application of the model. For example, items sensitivity to change can be extended to longer time intervals to better describe the item ability to sustain such sensitivity. Interactions with baseline levels of cognitive functioning can also be incorporated into the analysis, given suitable sample sizes for such analysis. This will allow the examination of item sensitivity to change conditioned upon varying stages of cognitive impairment. These analyses can be applied to a variety of scales measuring a broad range of cognitive and functional abilities to study their sensitivity to assess *meaningful* changes in individual and global neurocognitive scores over time. Future work is needed to explore further extensions of methodological approaches, such as the LFMC illustrated here, that will allow us to use more sensitive measures to detect fine-grained early pathological changes in cognition.

References

Aisen PS, Schneider LS, Sano M, Diaz-Arrastia R, van Dyck CH, Weiner MF, … &, Thal LJ. High-dose B vitamin supplementation and cognitive decline in Alzheimer disease: a randomized controlled trial. JAMA: Journal of the American Medical Association. 2008; 300:1774–1783. [PubMed: 18854539]

Akaike H. A new look at the statistical model identification. Automatic Control, IEEE Transactions on. 1974; 19:716–723.

Amieva H, Le Goff M, Millet X, Orgogozo JM, Pérès K, Barberger-Gateau P, … &, Dartigues JF. Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. Annals of Neurology. 2008; 64:492–498. [PubMed: 19067364]

Becker RE, Greig NH, Giacobini E. Why do so many drugs for Alzheimer's disease fail in development? Time for new methods and new practices? Journal of Alzheimer's Disease. 2008; 15:303–325.

Benge JF, Balsis S, Geraci L, Massman PJ, Doody RS. How well do the ADAS-cog and its subscales measure cognitive dysfunction in Alzheimer's disease? Dementia and Geriatric Cognitive Disorders. 2009; 28:63–69. [PubMed: 19641319]

Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin. 1980; 88:588–606.

Bollen, KA.; Curran, PJ. Latent curve models: A structural equation perspective. Hoboken, NJ: Wiley; 2006.

Bondi MW, Monsch AU, Galasko D, Butters N, Salmon DP, Delis DC. Preclinical cognitive markers of dementia of the Alzheimer type. Neuropsychology. 1994; 8:374–384.

Bondi MW, Salmon DP, Monsch AU, Galasko D, Butters N, Klauber MR, … &, Saitoh T. Episodic memory changes are associated with the APOE-epsilon 4 allele in nondemented older adults. Neurology. 1995; 45:2203–2206. [PubMed: 8848194]

Broich, K.; Schlosser-Weber, G.; Weiergräber, M.; Hampel, H. Regulatory Requirements on Clinical Trials in Alzheimer's Disease. In: Hampel, H.; Carrillo, MC., editors. Alzheimer's Disease – Modernizing Concept, Biological Diagnosis and Therapy: Vol. 28. Advances in Biological Psychiatry. Basel, Switzerland: Karger; 2012. p. 168-178.

Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociological Methods Research. 1992; 21:230–258.

Bussire T, Gold G, Kvari E, Giannakopoulos P, Bouras C, Perl DP, … &, Hof PR. Stereologic analysis of neurofibrillary tangle formation in prefrontal cortex area 9 in aging and Alzheimer's disease. Neuroscience. 2003; 117:577–592. [PubMed: 12617964]

Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, … &, van Belle G. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. Journal of Clinical Epidemiology. 2008; 61:1018–1027. [PubMed: 18455909]

De Jager C, Blackwell AD, Budge MM, Sahakian BJ. Predicting cognitive decline in healthy older adults. The American Journal of Geriatric Psychiatry. 2005; 13:735–740. [PubMed: 16085791]

Esiri MM, Chance SA. Cognitive reserve, cortical plasticity and resistance to Alzheimer's disease. Alzheimers Research & Therapy. 2012; 4:1–8.

Feldman HH, Doody RS, Kivipelto M, Sparks DL, Waters DD, Jones RW, … &, Breazna A. Randomized controlled trial of atorvastatin in mild to moderate Alzheimer disease LEADe. Neurology. 2010; 74:956–964. [PubMed: 20200346]

Flynn KE, Dombeck CB, DeWitt EM, Schulman KA, Weinfurt KP. Using item banks to construct measures of patient reported outcomes in clinical trials: investigator perceptions. Clinical Trials. 2008; 5:575–586. [PubMed: 19029206]

Folstein MF, Folstein SE, McHugh PR. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12:189–198. [PubMed: 1202204]

Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clinical and Experimental Rheumatology. 2005; 23:33–37.

Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research. 1992; 18:117–144. [PubMed: 1459160]

Kurz A, Farlow M, Quarg P, Spiegel R. Disease stage in Alzheimer disease and treatment effects of rivastigmine. Alzheimer Disease & Associated Disorders. 2004; 18:123–128. [PubMed: 15494617]

Llano DA, Laforet G, Devanarayan V. Derivation of a new ADAS-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to Alzheimer disease. Alzheimer Disease & Associated Disorders. 2011; 25:73–84. [PubMed: 20847637]

Lo RY, Jagust WJ, Aisen P, Jack CR, Toga AW, Beckett L, … &, Chowdhury M. Predicting missing biomarker data in a longitudinal study of Alzheimer disease. Neurology. 2012; 78:1376–1382. [PubMed: 22491869]

Marquis S, Moore MM, Howieson DB, Sexton G, Payami H, Kaye JA, Camicioli R. Independent predictors of cognitive decline in healthy elderly persons. Archives of Neurology. 2002; 59:601–606. [PubMed: 11939895]

McArdle, JJ. Five steps in the structural factor analysis of longitudinal data. In: Cudeck, R.; MacCallum, RC., editors. Factor analysis at 100: Historical developments and future directions. Mahwah, NJ: Lawrence Erlbaum Associates; 2007. p. 99-130.

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58:525–543.

Meredith, W.; Horn, J. The role of factorial invariance in modeling growth and change. In: Sayer, AG.; Collins, LM., editors. New Methods for the Analysis of Change. Washington, D.C: American Psychological Association; 2001. p. 201-240.

Millsap, RE.; Meredith, W. Factorial invariance: Historical perspectives and new problems. In: Cudeck, R.; MacCallum, RC., editors. 100 years of factor analysis. Mahwah, NJ: Lawrence Erlbaum Associates; 2007. p. 131-152.

Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, … &, Thai LJ. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's disease assessment scale that broaden its scope. Alzheimer Disease & Associated Disorders. 1997; 11:13–21. [PubMed: 9194962]

Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]

Mungas D, Reed BR. Application of item response theory for the development of a global functioning measure of dementia with linear measurement properties. Statistics in Medicine. 2000; 19:1631–1644. [PubMed: 10844724]

Mungas D, Reed BR, Kramer JH. Psychometrically matched measures of global cognition, memory, and executive function for the assessment of cognitive decline in older persons. Neuropsychology. 2003; 17:380–392. [PubMed: 12959504]

Muthén, LK.; Muthén, BO. Mplus User's Guide. 7. Los Angeles, CA: Muthén & Muthén; 1998–2012.

Proust-Lima C, Amieva H, Dartigues JF, Jacqmin-Gadda H. Sensitivity of Four Psychometric Tests to Measure Cognitive Changes in Brain Aging-Population–based Studies. American Journal of Epidemiology. 2007; 165:344–350. [PubMed: 17105962]

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2015. http://www.R-project.org

Rafii MS, Walsh S, Little JT, Behan K, Reynolds B, Ward C, … &, Aisen PS. A phase II trial of huperzine A in mild to moderate Alzheimer disease. Neurology. 2011; 76:1389–1394. [PubMed: 21502597]

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin. 1993; 114:552–566. [PubMed: 8272470]

Robert P, Ferris S, Gauthier S, Ihl R, Winblad B, Tennigkeit F. Review of Alzheimer's disease scales: Is there a need for a new multi-domain scale for therapy evaluation in medical practice. Alzheimer's Research & Therapy. 2010; 2:24, 1–13.

Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. The American Journal of Psychiatry. 1984; 141:1356–1364. [PubMed: 6496779]

Salmon DP, Lange KL. Cognitive screening and neuropsychological assessment in early Alzheimer's disease. Clinics in Geriatric Medicine. 2001; 17:229–254. [PubMed: 11375134]

Salsman JM, Lai JS, Hendrie HC, Butt Z, Zill N, Pilkonis PA, … &, Cella D. Assessing psychological well-being: self-report instruments for the NIH Toolbox. Quality of Life Research. 2014; 23:205–215. [PubMed: 23771709]

Sano M, Bell KL, Galasko D, Galvin JE, Thomas RG, van Dyck CH, Aisen PS. A randomized, double-blind, placebo-controlled trial of simvastatin to treat Alzheimer disease. Neurology. 2011; 77:556–563. [PubMed: 21795660]

Schaie KW, Maitland SB, Willis SL, Intrieri RC. Longitudinal invariance of adult psychometric ability factor structures across 7 years. Psychology and Aging. 1998; 13:8–20. [PubMed: 9533186]

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6:461–464.

Sclove LS. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika. 1987; 52:333–343.

Shah RC, Kamphuis PJ, Leurgans S, Swinkels SH, Sadowsky CH, Bongers A, … &, Bennett DA. The S-Connect study: results from a randomized, controlled trial of Souvenaid in mild-to-moderate Alzheimer's disease. Alzheimers Research & Therapy. 2013; 5:59, 1–9.

Skinner J, Carvalho JO, Potter GG, Thames A, Zelinski E, Crane PK. …Alzheimer's Disease Neuroimaging Initiative. The Alzheimer's disease assessment scale-cognitive-plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness in MCI. Brain Imaging and Behavior. 2012; 6:489–501. [PubMed: 22614326]

Slick, DJ. Psychometrics in neuropsychological assessment. In: Strauss, E.; Sherman, EM.; Spreen, O., editors. A compendium of neuropsychological tests: Administration, norms, and commentary. Oxford University Press; USA: 2006. p. 3-43.

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, … &, Phelps CH. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011; 7:280–292.

Steiger, JH.; Lind, JC. Statistically-based tests for the number of common factors. Paper presented at the Annual Spring Meeting of the Psychometric Society; Iowa City, IA. 1980 May.

Stevens, JP. Applied Multivariate Statistics for the Social Sciences. 5. New York, NY: Routeledge; 2009.

Suzuki T, Shimada H, Makizako H, Doi T, Yoshida D, Ito K, … &, Kato T. A randomized controlled trial of multicomponent exercise in older adults with mild cognitive impairment. PloS One. 2013; 8:e61483. [PubMed: 23585901]

Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika. 1973; 38:1–10.

Ueckert S, Plan EL, Ito K, Karlsson MO, Corrigan B, Hooker AC. Alzheimer's Disease Neuroimaging Initiative. Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. Pharmaceutical Research. 2014; 31:2152–2165. [PubMed: 24595495]

Welsh K, Butters N, Hughes J, Mohs R, Heyman A. Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. Archives of Neurology. 1991; 48:278–281. [PubMed: 2001185]

Welsh KA, Butters N, Hughes JP, Mohs RC, Heyman A. Detection and staging of dementia in Alzheimer's disease: Use of the neuropsychological measures developed for the Consortium to Establish a Registry for Alzheimer's Disease. Archives of Neurology. 1992; 49:448–452. [PubMed: 1580805]

West, SG.; Taylor, AB.; Wu, W. Model fit and model selection in structural equation models. In: Hoyle, RH., editor. Handbook of Structural Equation Models. New York, NY: Guilford; 2014. p. 209-231.

Wilson RS, Leurgans SE, Boyle PA, Bennett DA. Cognitive decline in prodromal Alzheimer disease and mild cognitive impairment. Archives of Neurology. 2011; 68:351–356. [PubMed: 21403020]

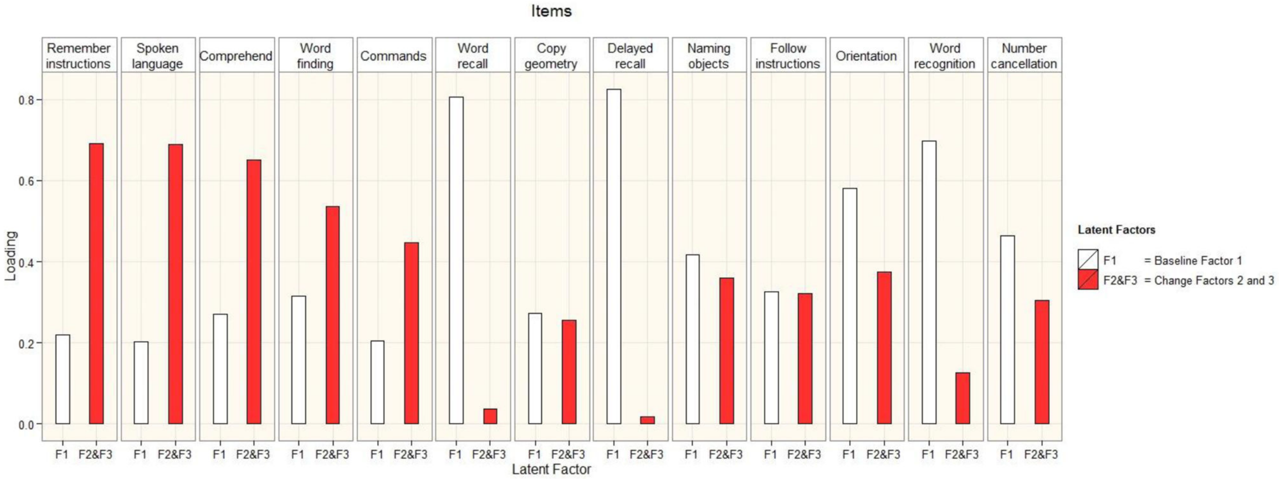

Figure 1.

Comparison of item loadings across baseline and change factors. The first five panels illustrate the loadings for the items showing greater sensitivity to change. The other panels compare the loadings for the remaining 8 items in the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog).

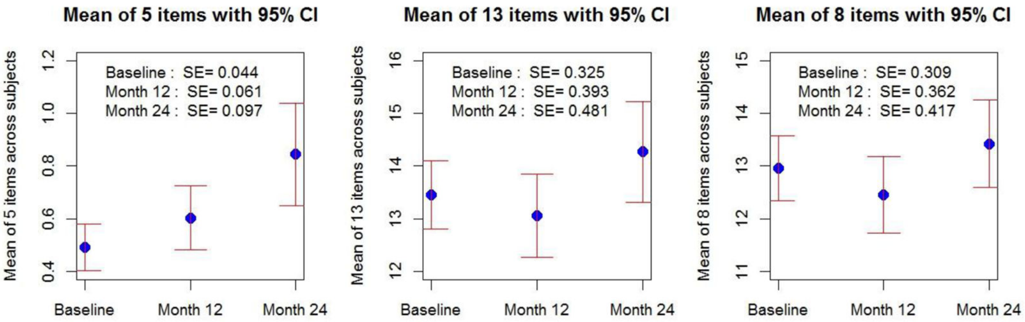

Figure 2.

Mean score trend across time by item group with 95% confidence intervals (CI) and corresponding standard errors (SEs).

Table 1

Sample Characteristics at Baseline by Diagnostic Group

Characteristic	Total *(n=1,217)*		CN *(n=343)*		MCI *(n=764)*		AD *(n=110)*	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Demographic								
Age (years)	73.44	7.08	74.21	5.69	72.83	7.47	75.20	7.75
Education (years)	16.11	2.75	16.44	2.63	16.02	2.81	15.69	2.61
Sex								
Male (%)	56.50%		51.60%		58.60%		57.30%	
Female (%)	43.50%		48.40%		41.40%		42.70%	
Global cognition and function (baseline)								
MMSE	27.30	2.50	29.06	1.15	27.58	1.81	23.19	2.05
CDR-SB	1.52	1.63	.04	.18	1.51	.88	4.37	1.63
ADAS-Cog-13	16.50	8.89	9.43	4.18	16.63	6.83	28.85	7.62

Note. CN = cognitively normal; MCI = mild cognitive impairment; AD = Alzheimer’s disease; MMSE = Mini Mental State Examination; CDR-SB = Clinical Dementia Rating-Sum of Boxes; ADAS-Cog13 = Alzheimer’s disease Assessment Scale-13-item Cognitive Subscale.

Table 2

ADAS-Cog-13 Items Submitted to Analysis

Item Number in the ADNI Database	Cognitive Domain	Item Description	Range
Item 1	Memory	Word-recall-Trials 1 to 3[a]	0–10
Item 2	Language	Following commands	0–5
Item 3	Constructional Praxis	Copy geometric forms	0–5
Item 4	Memory	Delayed word recall	0–10
Item 5	Language	Naming objects	0–5
Item 6	Ideational Praxis	Following instructions[b]	0–5
Item 7	Orientation	Orientation	0–8
Item 8	Memory	Word recognition	0–12
Item 9	Memory	Remember instructions[b]	0–5
Item 10	Language	Comprehension of spoken language[b]	0–5
Item 11	Language	Word finding difficulty[b]	0–5
Item 12	Language	Spoken language ability[b]	0–5
Item 13	Attention/Executive Function	Number cancellation	0–40

Notes. ADAS-Cog13 = Alzheimer’s disease Assessment Scale-13-item Cognitive Subscale; ADNI = Alzheimer’s Disease Neuroimaging Study.

Items in the ADAS-Cog are also referred to as “tasks.”

[a] Range of 0–10 per trial.

[b] Clinician rated item.

Table 3

Fit Indices for the Models Testing Measurement Invariance

Models	AIC	BIC	SABIC	χ^2	*df*	*p*	CFI	TLI	RMSEA	90% CI LL and UL
Constant Factor Model	63855.31	64442.29	64077.00	2183.83	704	< 0.001	0.85	0.84	0.04	[0.040, 0.044]
Baseline and Change Model[a]	62663.19	63316.52	62909.94	1513.63	691	< 0.001	0.92	0.91	0.03	[0.029, 0.033]
Varying Factor Model	62644.90	63359.48	62914.79	1544.70	679	< 0.001	0.91	0.90	0.03	[0.030, 0.035]

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean squared error of approximation.

[a] Model selected.

Table 4

Standardized Loadings and Standard Errors per Estimated Factor

		Baseline Factor		Change Factors 2 and 3	
Item	Item Description	Estimate	SE	Estimate	SE
Item 1	Word recall-Trials 1 to 3	.81	.01	.04	.00
Item 2	Commands	.20	.04	.45	.08
Item 3	Copy geometric forms	.27	.02	.26	.08
Item 4	Delayed word recall	.83	.01	.02	.02
Item 5	Naming objects	.42	.03	.36	.07
Item 6	Following instructions	.33	.03	.32	.08
Item 7	Orientation	.58	.03	.37	.05
Item 8	Word recognition	.70	.02	.13	.03
Item 9	Remember instructions	.22	.05	.69	.06
Item 10	Comprehension	.27	.05	.65	.06
Item 11	Word finding	.32	.04	.54	.04
Item 12	Spoken language	.20	.04	.69	.05
Item 13	Number cancellation	.46	.03	.30	.04

Table 5

Ranking of Items on Two Different Latent Factors Across Time

Baseline Factor			Change Factors 2 and 3		
Item	**Label**	**Loading (ranked from highest to lowest)**	**Item**	**Label**	**Loading (ranked from highest to lowest)**
Item 4	Delayed word recall	.83	Item 9	Remember instructions	.69
Item 1	Word recall-Trials 1 to 3	.81	Item 12	Spoken language	.70
Item 8	Word recognition	.70	Item 10	Comprehension	.65
Item 7	Orientation	.58	Item 11	Word finding	.54
Item 13	Number cancellation	.46	Item 2	Commands	.45
Item 5	Naming objects	.42	Item 7	Orientation	.37
Item 6	Following instructions	.33	Item 5	Naming objects	.36
Item 11	Word finding	.32	Item 6	Following instructions	.32
Item 3	Copy geometric forms	.27	Item 13	Number cancellation	.30
Item 10	Comprehension	.27	Item 3	Copy geometric forms	.26
Item 9	Remember instructions	.22	Item 8	Word recognition	.13
Item 2	Commands	.20	Item 1	Word recall-Trials 1 to 3	.04
Item 12	Spoken language	.20	Item 4	Delayed word recall	.02

Table 6

Mean Observed Score by Item Group, Diagnostic Group, and Time

	5 items		**8 items**		**13 items**	
Time	**M**	**SD**	**M**	**SD**	**M**	**SD**
			Total			
Baseline	.49	.97	12.96	6.84	13.45	7.20
Month 12	.60	1.36	12.45	8.02	13.06	8.70
Month 24	.85	2.15	13.42	9.23	14.27	10.67
			CN			
Baseline	.22	.61	8.92	4.17	9.14	4.32
Month 12	.22	.59	7.86	4.35	8.09	4.53
Month 24	.21	.60	7.95	4.23	8.16	4.46
			MCI			
Baseline	.52	.90	13.64	6.08	14.17	6.29
Month 12	.62	1.25	13.08	7.12	13.70	7.64
Month 24	.80	1.71	14.23	8.24	15.03	9.18
			AD			
Baseline	1.57	2.01	26.71	5.74	28.29	6.46
Month 12	2.39	3.02	29.64	7.06	32.04	8.43
Month 24	4.71	5.50	33.29	8.96	38.00	13.49

Note. M = mean; SD = standard deviation; CN = cognitively normal; MC I= mild cognitive impairment; AD = Alzheimer’s disease.

Table 7

Standardized Slope Parameter Estimates for the Unconditional LGCM Model by Item Group

Item Group	Mean Slope Estimate	SE	*p*	BIC
5-Item	.24	.06	< 0.001	4583.27
8-Item	.16	.08	0.037	8752.61
13-Item	.20	.06	0.002	9014.26

Note. LGCM = Latent growth curve model; SE = Standard error; BIC = Bayesian Information Criterion.