Citation: CPT Pharmacometrics Syst. Pharmacol. (2017) 6, 635–641; doi:10.1002/psp4.12219

ORIGINAL ARTICLE

A Longitudinal Item Response Theory Model to Characterize Cognition Over Time in Elderly Subjects

Marc Vandemeulebroecke[1]*, Björn Bornkamp[1], Tillmann Krahnke[2], Johanna Mielke[1], Andreas Monsch[3] and Peter Quarg[1]

For drug development in neurodegenerative diseases such as Alzheimer's disease, it is important to understand which cognitive domains carry the most information on the earliest signs of cognitive decline, and which subject characteristics are associated with a faster decline. A longitudinal Item Response Theory (IRT) model was developed for the Basel Study on the Elderly, in which the Consortium to Establish a Registry for Alzheimer's Disease – Neuropsychological Assessment Battery (with additions) and the California Verbal Learning Test were measured on 1,750 elderly subjects for up to 13.9 years. The model jointly captured the multifaceted nature of cognition and its longitudinal trajectory. The word list learning and delayed recall tasks carried the most information. Greater age at baseline, fewer years of education, and positive APOEε4 carrier status were associated with a faster cognitive decline. Longitudinal IRT modeling is a powerful approach for progressive diseases with multifaceted endpoints.

CPT Pharmacometrics Syst. Pharmacol. (2017) **6**, 635–641; doi:10.1002/psp4.12219; published online 23 June 2017.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Cognition in the elderly has previously been modeled by longitudinal disease progression models and a cross-sectional Item Response Theory (IRT) model, across a spectrum of healthy subjects and those with Alzheimer's disease.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ A longitudinal IRT model was developed, jointly capturing the multifaceted nature of cognition and its longitudinal trajectory, to investigate which cognitive domains carry the most information on the earliest signs of cognitive decline, and which subject characteristics are associated with a faster decline, in generally healthy elderly subjects.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

☑ Verbal episodic memory as assessed by word list learning and delayed recall tasks carried the most information on early signs of cognitive decline. Greater age at baseline, fewer years of education, and positive APOEε4 carrier status were associated with a faster cognitive decline.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

☑ Focusing on verbal episodic memory may increase our chances to detect putative drug effects in the early stages of cognitive decline, in neurodegenerative diseases such as Alzheimer's disease or otherwise. Longitudinal IRT modeling is a powerful approach for progressive diseases with multifaceted endpoints.

Understanding cognition in the elderly is key to drug development for the treatment of major neurodegenerative diseases. In 2015, dementia affected more than 46 million people worldwide, causing an estimated total cost of 818 billion USD.[1] Alzheimer's disease (AD) in particular, as the most common cause of dementia, is progressive and fatal, and the only existing treatments are symptomatic and of limited effectiveness. Today it is generally believed that the most promising approach to treat this devastating disease effectively is to intervene very early in its underlying pathological cascade.[2] It is therefore important to enhance our understanding of cognition in the elderly, its trajectory over time, and the earliest signs of cognitive decline. This can provide vital information for clinical trial design: emphasis can be put on those endpoints that are the most sensitive in the early stages; natural disease progression can serve as an epidemiological benchmark; and knowledge of major covariates can inform strategies of population enrichment or stratification. In this context, the goal of the present work was to contribute to understanding:

- which cognitive domains carry the most information on early signs of cognitive decline in the elderly; and
- which subject characteristics are associated with a faster cognitive decline over time.

Several models have been developed to capture cognition in the elderly, across a spectrum of cognitively healthy up to moderately impaired subjects. Many of these models draw data from the Alzheimer's Disease Neuroimaging Initiative[3] (ADNI) and focus on the longitudinal trajectory of an overall summary measure of cognition, often the Alzheimer's

[1]Novartis Pharma AG, Basel, Switzerland; [2]Freelance statistical consultant; [3]University Center for Medicine of Aging, Basel, Switzerland. *Correspondence: Marc Vandemeulebroecke (marc.vandemeulebroecke@novartis.com)
Received 14 November 2016; accepted 6 June 2017; published online on 23 June 2017. doi:10.1002/psp4.12219

Disease Assessment Scale – Cognitive Subscale (ADAS-Cog) or the Clinical Dementia Rating Scale – Sum of Boxes (CDR-SoB). Landmark models of this type include those by Ito *et al.*,[4,5] Rogers *et al.*,[6] Samtani *et al.*,[7,8] and Delor *et al.*[9] However, "cognition" is a complex notion, and any model for a "lump sum" summary measure can only be as good as the summary measure that it uses. Ueckert *et al.*[10] drew more information from the multifaceted nature of cognition by applying Item Response Theory[11] (IRT) to the ADAS-Cog at the individual test item level. IRT, with conceptual ideas dating back to the 1950s and 1960s (e.g., Rasch[12]), is a collection of latent variable models that has grown popular in psychometrics since the 1980s (e.g., through works by Lord[13] and Hambleton *et al.*[14]), and more recently also in pharmacometrics.[10,15,16] Among other advantages, it preserves each test item's information and does not collapse all information into a single summary score. The model by Ueckert *et al.*[10] was fundamentally a cross-sectional model that was built using combined baseline data from ADNI and seven clinical trials. The individual test items were related to a hypothesized underlying latent "disability" variable through a set of link functions (called "item characteristic functions" in IRT), the parameters of which were estimated from the data. The pattern of longitudinal progression was then investigated in a second step, by developing a linear progression model for the latent variable, using longitudinal data from another (separate) study, and fixing all parameters of the IRT model to their previously estimated values. A necessary assumption for this approach is that the link between the underlying disability and the measured test items, as obtained from baseline data only, applies universally at any timepoint.

The present article draws ideas from these works and adds to their perspectives, in terms of results as well as methodology. Our data source is a noninterventional longitudinal study of mostly cognitively healthy elderly subjects. We propose to consider both aspects, the multifaceted nature of cognition and its longitudinal development, jointly rather than separately. Instead of fitting an IRT model on cross-sectional baseline data and then applying it to longitudinal data, we develop a single *longitudinal* IRT model that takes into account all data at once. All parameters of the model, that is, those relating the test items to the latent variable and those describing the longitudinal decline, are estimated jointly based on all data. We compute the information content of each test item, and we investigate which baseline covariates are associated with a faster decline of the latent variable. By implementing the model in a fully Bayesian framework using three different software platforms, we obtain an informal comparison of their performance and ease of use for a complex statistical model.

METHODS

Available data

This work uses data from the Basel Study on the Elderly[17,18] (BASEL study), an observational study conducted by the Memory Clinic of the University Center for Medicine of Aging, Basel, Switzerland. The objective of this study was to identify presymptomatic markers of dementia in a group of previously healthy individuals. A cohort of elderly subjects were assessed by two neuropsychological test batteries: the Consortium to Establish a Registry for Alzheimer's Disease – Neuropsychological Assessment Battery[19] (CERAD-NAB) with its additions of Phonemic fluency[20] (S-words) and the Trail Making Test,[21] together constituting the CERAD-NAB-Plus, as well as by the California Verbal Learning Test[22] (CVLT). A subset of the initial cohort continued into a longitudinal observation phase, during which these assessments were biannually repeated. The study was approved by the local Ethics Committee and conducted in accordance with the Declaration of Helsinki;[23] all subjects provided informed consent. **Table 1** displays the 14 main test items of the two neuropsychological test batteries; ancillary measures were discarded. Word list learning and recall tasks had been purposely included from both batteries, in order to assess whether one performed better than the other.

Data cleaning was performed by the Memory Clinic, with minor additional cleaning by the authors (removing four spurious observations: one without subject ID, two with a

Table 1 Neuropsychological test items in the BASEL study

CERAD-NAB
Semantic fluency (animals)
— Task: Naming animals; Outcome: Number of words
Boston Naming Test
— Task: Recognize pictures; Outcome: Number of correct answers
Mini-Mental Status Examination (MMSE)
— Task: Respond to a range of cognitive tasks; Outcome: Score
Word List Learning
— Task: Learn words and recall them; Outcome: Number of correct words
Word List Delayed Recall
— Task: Recall those words again after delay; Outcome: Number of correct words
Word List Recognition
— Task: Identify those words in a list including distractor words; Outcome: Discriminability score
Constructional Praxis
— Task: Copy pictures; Outcome: Score (how many graphical aspects were correctly drawn)
Constructional Praxis Delayed Recall
— Task: Draw those pictures again after delay; Outcome: Score (how many graphical aspects were correctly drawn)
CERAD-NAB additions
Phonemic Fluency (S-words)
— Task: Name words starting with "S"; Outcome: Number of words
Trail Making Test (Part A)
— Task: Connect numbers; Outcome: Score (time needed)
Trail Making Test (Part B)
— Task: Connect numbers/letters; Outcome: Score (time needed)
California Verbal Learning Test (CVLT)
CVLT – Word List Learning
— Task: Learn words and recall them; Outcome: Number of correct words
CVLT – Word List Recognition
— Task: Identify those words in a list including distractor words; Outcome: Discriminability score
CVLT – Word List Long Delay Free Recall
— Task: Recall those words again after long delay; Outcome: Number of correct words

negative timepoint, and one empty observation; and using the earlier date in case of duplicate assessment dates for three subjects). The final dataset consisted of 52,370 neuropsychological test item values from 1,750 subjects. After inversion of the scale of the Trail Making Test (A and B), high values indicated high abilities on all test items. The available baseline covariates of interest were gender, age at baseline, Mini-Mental Status Examination (MMSE) at baseline, years of education, and APOEε4 genotype (a genetic risk factor for AD[24,25]).

Model development

The dual goal of identifying informative test items and assessing cognitive decline led to the choice of a modeling strategy that would combine IRT with longitudinal progression in a single model using all item-level data jointly. Model development began with descriptive and graphical data explorations to reveal the actual visit structure, baseline characteristics, and endpoint distribution and progression. In light of the different empirical distributions of the endpoints, the graded response model[26,27] for ordered categorical responses was chosen as a generic IRT building block to relate each test item to a common latent "ability" trait, for its flexibility and for parsimony. We assumed unidimensionality of the latent ability, which implies local independence of the test items (i.e., independence after conditioning on the subjects' abilities[13,28]). For computational feasibility, the raw item responses were condensed into fewer categories. The longitudinal progression was captured within the same model, using a linear progression pattern of the latent ability over time.

We denote the response category of subject s in test item j at time t by $Y_{s,t,j}$, the corresponding latent ability by $\theta_{s,t}$ (the same for any item), and we write $p_{s,t,j,k} = P(Y_{s,t,j} \leq k \mid \theta_{s,t})$. With each test item's raw responses condensed into K response categories of equal width, the graded response IRT model for the BASEL dataset is specified by $p_{s,t,j,K} = 1$ and

$$\text{logit}(p_{s,t,j,k}) = \log\left(\frac{p_{s,t,j,k}}{1-p_{s,t,j,k}}\right) = \kappa_{j,k} - \alpha_j \theta_{s,t},\ k = 1, \ldots, K-1.$$

The parameters α_j determine the steepness of the item characteristic functions and are called "discrimination parameters" in IRT. The parameters $\kappa_{j,k}$, known as "difficulty parameters", determine the difficulty of each response category of each item. Both do not depend on subject or time. After trying several values for K, we settled on $K = 9$ as a practical compromise between computational tractability and loss of information.

The longitudinal aspect is captured within the same model by setting

$$\theta_{s,t} = \gamma_{0,s} + \gamma_{1,s} \times t,$$

that is, by a linear progression of the latent ability with subject-specific intercepts and slopes. The slope parameters are modeled as $\gamma_{1,s} = \gamma^*_{1,s} + \mathbf{x}'_s \boldsymbol{\beta}$, where $\mathbf{x}_s$ is a subject-specific vector of covariates (gender, age at baseline, MMSE at baseline, years of education, APOEε4 carrier, and APOEε4 homozygous carrier), and $\boldsymbol{\beta}$ is a vector of coefficients. Continuous covariates were zero-centered and standardized for model fitting.

We cast the model in a Bayesian framework. As the use of flat priors is not advisable in nonlinear models like ours,[29] weakly informative prior distributions were used for all parameters except $\gamma_{0,s}$. "Weakly informative" here refers to the fact that the prior distributions had a wide spread compared to the likelihood function so that they did not have a major impact on the posterior distributions but stabilized the model fit. Independent $N(0,1)$ prior distributions were used for the subject-specific intercept parameters $\gamma_{0,s}$. This is an arbitrary calibration of the model, required to ensure identification of the discrimination and difficulty parameters. For the intercept $\gamma^*_{1,s}$ of the subject-specific slope parameters, a hierarchical prior was chosen to allow sharing of information between subjects; we denote its mean by μ. See the **Supplementary Material** for a full specification of the prior distributions.

Three elements are of primary interest in this model: the discrimination parameter and the Fisher information for each test item (high values indicate the most discriminatory and informative test items), and the regression coefficients for the individual slopes (to investigate which subject characteristics are associated with a steeper decline of cognitive ability). The **Supplementary Material** provides more detail on the computation of the Fisher information, which is also known as "item information" in IRT.

The model was implemented with Markov Chain Monte Carlo (MCMC) methods in three programming languages, to ensure reliability of results: WinBUGS 1.4.1,[30] JAGS 4.2.0,[31] and Stan 2.14.1.[32] All program runs were performed on a high-performance computing cluster. Pre- and postprocessing of all analyses were done in R 3.2.3.[33]

Model qualification

Convergence of the MCMC results was carefully checked in the different software implementations. MCMC trace plots were examined, the Gelman-Rubin diagnostic[34] was checked, and effective sample sizes were computed.[35] Assumptions specific to the IRT model, such as unidimensionality of the latent ability and invariance of the item and ability parameters, were checked following Hambleton *et al.*[14] Visual predictive checks were performed, including a detailed categorical version following Ueckert *et al.*,[10] in which the observed fraction of subjects scoring in each category of any item was compared against the corresponding model-simulated confidence band over time.

RESULTS

Available data

Table 2 shows baseline characteristics of the total BASEL study as used in this analysis. The cohort consisted of 1,750 elderly, well-educated, mostly cognitively healthy subjects with a normal representation of APOEε4 genotypes[36] and twice as many males as females. A subset of 718 subjects (41.0%) provided data on more than one visit. This subset had been determined by selecting the APOEε4 carriers and approximately twice as many age-, education-,

Table 2 Baseline characteristics of the BASEL study population

Number of subjects	**1750**	
Gender		
Female	598	(34.2%)
Male	1152	(65.8%)
APOEϵ4 genotype		
Noncarrier	1234	(70.5%)
Heterozygous carrier	347	(19.8%)
Homozygous carrier	23	(1.3%)
Missing information	146	(8.3%)
Age at baseline [years]		
Mean (SD)	69.9	(8.0)
(min, max)	(49, 92)	
MMSE at baseline		
Mean (SD)	28.6	(1.5)
(min, max)	(17, 30)	
Missing information	4	(0.2%)
Years of education		
Mean (SD)	12.6	(3.3)
(min, max)	(4, 43)	

The maximum years of education would be 23 disregarding two extreme values; mean and SD would not change meaningfully. SD, standard deviation.

and gender-matched noncarriers from the initial cohort.[17] With individual durations of observation time of up to 13.9 years, this longitudinal subset accounted for 82.3% of the total number of observations. The CERAD-NAB additions and the CVLT were recorded on the same day, but often on a different day than the CERAD-NAB: for 10% of the subject-visits the time difference was greater than 17 weeks, with extreme differences of up to 185 weeks. Subjects without complete covariate information were excluded from the analysis; the reduced dataset consisted of 1,604 subjects and 50,917 test item values. A sensitivity analysis was performed on all subjects using multiple imputation for missing covariates (see **Supplementary Material**).

Longitudinal IRT model

The results of the JAGS and Stan implementations were identical (up to sampling error), and they are described in the remainder of this section. **Table 3** and **Figure 1** show the posterior distributions of the discrimination parameters. The word list learning and delayed recall tasks of the CERAD-NAB and the CVLT were the most discriminative (greatest discrimination parameter), i.e., the most sensitive to subtle differences in cognitive ability. The posterior estimates of the difficulty parameters are included in the **Supplementary Material,** as well as the item characteristic functions of each test item.

Figure 2 displays the item information. It provides an impression on which tests carry the most information (i.e., are sensitive to differences in ability), and also (unlike the discrimination parameters) where on the scale of underlying abilities they are most informative. Again, the word list learning and delayed recall tasks of both test batteries were the most informative over a broad range of abilities. The word list learning tasks showed some advantage at very high abilities (such as $\theta>2$). The MMSE and the word list recognition tasks were informative only in the low ability range. Semantic fluency, the Boston Naming Test, Constructional praxis, Constructional praxis delayed, Phonemic fluency, and the Trail Making Test (parts A and B) carried little to almost no information in the BASEL sample. The relative contribution of each item to the total information, calculated as the area under the item information curve divided by the sum of all such areas, is annotated inside each panel. Together, the word list learning and delayed recall tasks carried almost half (47.1%) of the total information (not accounting for their quasiduplication). Among these, the items of the CVLT carried slightly more information than their respective counterparts of the CERAD-NAB.

Table 3 Posterior means and distributions of the discrimination parameters

	Mean	**SD**	**Q2.5**	**Q97.5**
Semantic fluency	1.25	0.04	1.16	1.34
Boston Naming Test	1.04	0.05	0.94	1.14
MMSE	1.77	0.09	1.59	1.96
Word list learning	2.05	0.06	1.93	2.18
Word list delayed recall	2.08	0.06	1.96	2.21
Word list recognition	1.79	0.10	1.59	1.99
Constructional praxis	0.79	0.05	0.69	0.89
Constr. praxis delayed	1.06	0.04	0.97	1.15
Phonemic fluency	1.14	0.05	1.05	1.23
Trail Making Test (A)	1.08	0.05	0.99	1.17
Trail Making Test (B)	1.60	0.06	1.49	1.72
CVLT: Learning	2.42	0.09	2.25	2.60
CVLT: Recognition	1.89	0.11	1.68	2.12
CVLT: Delayed	2.28	0.09	2.11	2.45

SD, standard deviation. Q2.5, 2.5% quantile. Q97.5, 97.5% quantile.

The posterior mean of the typical slope parameter μ was -0.013 (2.5% and 97.5% quantiles [–0.025, -0.001]),

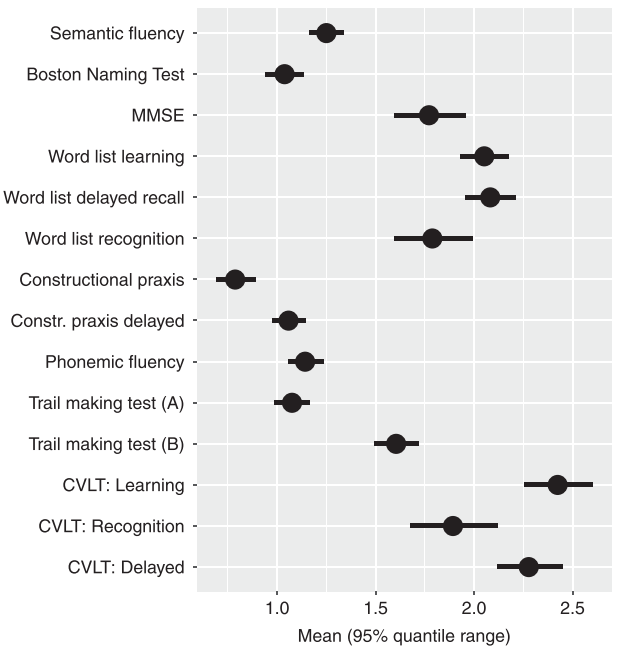

Figure 1 Posterior means and 95% quantile ranges of the discrimination parameters.

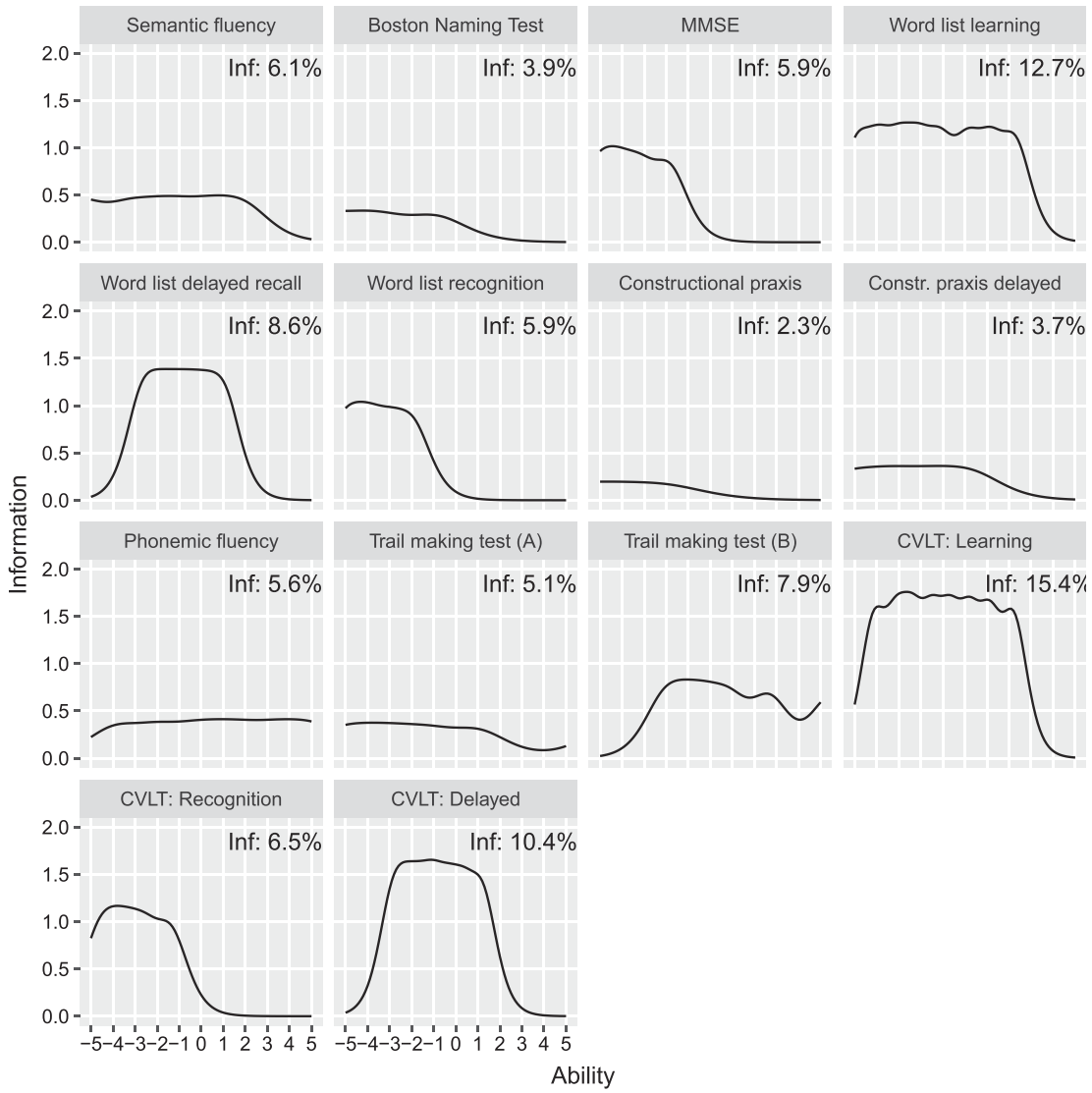

Figure 2 Item information curves. Inf, Relative contribution of each item to the total information.

indicating that the latent ability declined by 0.013 units per year for a "typical" subject. **Table 4** displays the regression coefficients of the baseline covariates for the individual slopes. Age at baseline, years of education, and APOEε4 carrier status showed a significant association with the slopes; gender and MMSE at baseline did not. Due to the arbitrary calibration of the latent ability, the numbers can only be interpreted in relative terms: older subjects, subjects with fewer years of education, and APOEε4 carriers (particularly homozygous carriers) showed a faster declining ability.

Model qualification

The model implementation in WinBUGS was inefficient and unstable. The posterior distributions of the IRT parameters reached effective sample sizes of only 2–36 before the MCMC process ran into numerical errors. The JAGS and Stan implementations were stable and more efficient. Using 10 chains for each with 5,000 iterations after burn-in, they were numerically stable and yielded acceptable Gelman–Rubin convergence diagnostics. Run times were similar for Stan and JAGS, but Stan produced more efficient chains:

Table 4 Regression coefficients for the individual slopes

	Mean	SD	*P*
Gender	−0.003	0.008	0.345
Age at baseline	−0.049	0.004	< 0.001
MMSE at baseline	0.006	0.005	0.128
Years of education	0.008	0.004	0.027
APOEϵ4 carrier	−0.019	0.008	0.008
APOEϵ4 homozygous carrier	−0.038	0.023	0.051

The binary covariates were coded as gender [1: female, 0: male], APOEϵ4 carrier [1: yes, 0: no], and APOEϵ4 homozygous carrier [1: yes, 0: no]. The covariates age at baseline, MMSE at baseline and years of education were zero-centered and standardized for model fitting. Mean, posterior mean. SD, posterior standard deviation. $P = 1 - \Phi(|\text{Mean}|/\text{SD})$, where Φ is the cumulative distribution function of the standard normal distribution, approximately corresponding to a two-sided *P*-value in a non-Bayesian setting.

the mean effective sample size across the non-subject-specific parameters was 506 for JAGS and 3,791 for Stan, based on the 5,000 inference samples each. The fact that the implementations in JAGS and Stan provided identical results shows the reliability of the results.

Assumptions specific to the IRT model were verified (not shown here). Visual Predictive Checks showed good agreement between the observed data and simulations from the model, even at the granular categorical level. The **Supplementary Material** provides more details on the model qualification and the performance comparison between JAGS and Stan.

DISCUSSION

Based on the BASEL study, in which 14 neuropsychological tests items were assessed on 1,750 elderly subjects for up to 13.9 years, the goals of this work were to investigate which cognitive domains carry the most information on the earliest signs of cognitive decline in the elderly, and which subject characteristics are associated with a faster cognitive decline. Answering these questions, and more generally developing a suitable methodology for answering them, is relevant for drug development against neurodegenerative diseases such as AD, where a "disease-modifying" early intervention is believed to be key. We developed a fully longitudinal IRT model that allowed capturing the multifaceted nature of cognition and its longitudinal development within one and the same model, based on all item-level data at once. With this, we combined ideas of the "giants on whose shoulders we stand," such as Ito *et al.*[4,5] and others for the longitudinal progression, and Ueckert *et al.*[10] for IRT.

A graded response model[26] was chosen for the link between latent ability and observable test items, while the longitudinal trajectory of the latent ability was modeled by a linear function over time. The association of subject-specific covariates with the slope of progression was embedded in the same model. The modeling approach allowed using the exact actual dates for each separate test item (per subject and visit), which was important, since different items of the same nominal visit were sometimes measured very far apart. The model was implemented in a Bayesian framework in three programming languages.

We found that the word list learning and delayed recall tasks of the CERAD-NAB and the CVLT were the most sensitive and carried the most information in the BASEL study, over a broad range of abilities. Among these, the items of the CVLT carried slightly more information than their respective counterparts of the CERAD-NAB. The MMSE and the word list recognition tasks were informative only in the low ability range, and the other test items carried little to almost no information. Greater age at baseline, fewer years of education, and positive APOEε4 carrier status were associated with a faster cognitive decline, with an even faster decline for homozygous carriers. Among the three software implementations, JAGS and Stan were stable and provided identical results, but Stan was more efficient.

Although the BASEL study population is predominantly cognitively healthy, our results confirm earlier findings in cognitively impaired or mixed populations. For example, Ueckert *et al.*[10] found that a delayed word recall task carried the most information in a population with Mild Cognitive Impairment (MCI), followed by word recall, orientation, and word recognition. Beck *et al.*[37] noted that verbal episodic memory tasks of the CVLT were more sensitive than corresponding tasks of the CERAD-NAB. Baseline age and APOEε4 carrier status were also associated with the slope of progression in the analyses by Ito et al.[5] (in a mixture of healthy subjects and those with MCI and mostly mild AD), Samtani *et al.*[7] (mostly mild AD), and Rogers *et al.*[6] (mild and moderate AD). While we excluded subjects with missing covariate information from our analysis, a sensitivity analysis on all subjects using multiple imputation for missing covariates showed very similar results (see **Supplementary Material**). Due to a lack of data, we could not investigate the influence of laboratory or imaging markers.

We fitted our model on all subjects (with nonmissing covariates), rather than only on the subset that provided longitudinal data, in order to obtain maximal information on the IRT parameters. The Bayesian methodology allowed the longitudinal slope estimates to remain vague for subjects with little (or no) follow-up time. Recent disease progression models for AD have used nonlinear progression patterns[6–8,38] and/or the concept of disease onset time.[9,38] Our choice of a linear progression pattern may be considered appropriate for the BASEL study population, which displays only small cognitive changes over time. It is in line with earlier models[4,5] and ensured computational tractability when embedding it into an IRT model. The assumption that a single dominant latent trait influences all observable test items was checked successfully; it was therefore not necessary to consider multidimensional link models (which quickly reach the limits of tractability[39]). It should be noted that one of the test items in the BASEL study, the MMSE, is derived from various cognitive tasks. Because of its widespread standard use, we still treated it as a single-entity item, acknowledging that building its summary score results in a loss of information. Another potential limitation of our work is that we condensed the raw item responses into nine response categories for each item (including the Trail Making Tests A and B whose raw values are quasicontinuous), and this impacts the information provided by these items. However, this choice greatly reduced the complexity of the model, and the results appear robust. Sensitivity analyses with more response categories required longer run times but yielded the same conclusions.

In summary, verbal episodic memory as assessed by word list learning and delayed recall tasks may be a promising domain for the detection of early signs of cognitive decline in the elderly. Longitudinal IRT modeling, as applied here in a mostly healthy elderly population, is a suitable method to capture the multifaceted nature of cognition and its longitudinal trajectory jointly. It is computationally more intensive than cross-sectional IRT models, but it allows the estimation of the IRT parameters based on all data. It would be of interest to apply this method also to a cohort with prodromal or mild AD, in this case possibly with a

nonlinear progression pattern. In general, longitudinal IRT modeling appears to be a powerful approach for progressive diseases with multifaceted endpoints.

Acknowledgments. The authors thank Heinz Schmidli for helpful advice, and the two anonymous reviewers whose comments helped to improve the original article.

Conflict of Interest/Disclosure. B.B., J.M., P.Q., and M.V. are employed by Novartis Pharma AG, Basel, Switzerland, which co-sponsored the BASEL study. T.K. was employed by Novartis Pharma AG, Basel, Switzerland when this work was conducted; he now works as a freelance statistical consultant. A.M. is the Principal Investigator of the BASEL study.

Author Contributions. M.V., P.Q. and T.K. designed the analysis; J.M., M.V. and B.B. performed it. M.V. wrote the manuscript; B.B., T.K., J.M. and P.Q. reviewed it. A.M. is the Principal Investigator of the BASEL study.

1. World Alzheimer Report. The global impact of dementia. <http://www.alz.co.uk/research/world-report-2015> (2015).
2. Jack C.R., *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9**, 119–128 (2010).
3. Weiner, M.W., *et al.* The Alzheimers Disease Neuroimaging Initiative: progress report and future plans. *Alzheimers Dement.* **6**, 202–211 (2010).
4. Ito K., Ahadieh S., Corrigan B., French J., Fullerton T. & Tensfeldt T. Disease progression meta-analysis model in Alzheimer's disease. *Alzheimers Dement.* **6**, 39–53 (2010).
5. Ito K., *et al.* Disease progression model for cognitive deterioration from Alzheimer's Disease Neuroimaging Initiative database. *Alzheimers Dement.* **7**, 151–160 (2011).
6. Rogers J.A., *et al.* Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta-regression meta-analysis. *J. Pharmacokinet. Pharmacodynam.* **39**, 479–498 (2012).
7. Samtani M.N., *et al.* An improved model for disease progression in patients from the Alzheimer's Disease Neuroimaging Initiative. *J. Clin. Pharmacol.* **52**, 629–644 (2012).
8. Samtani M.N., *et al.* Disease progression model in subjects with mild cognitive impairment from the Alzheimer's Disease Neuroimaging Initiative: CSF biomarkers predict population subtypes. *Br. J. Clin. Pharmacol.* **75**, 146–161 (2013).
9. Delor I., Charoin J.-E., Gieschke R., Retout S. & Jacqmin P. Modeling Alzheimer's disease progression using disease onset time and disease trajectory concepts applied to CDR-SoB scores from ADNI. *CPT Pharmacometrics Syst. Pharmacol.* **2**, e78 (2013).
10. Ueckert S., Plan E.L., Ito K., Karlsson M.O., Corrigan B.W. & Hooker A.C. Improved utilization of ADAS-Cog assessment data through item response theory based pharmacometric modeling. *Pharm. Res.* **31**, 2152–2165 (2014).
11. DeMars C. Item Response Theory (Oxford University Press, Oxford, UK; 2010).
12. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Copenhagen, 1960). Expanded edition (University of Chicago Press, Chicago, IL, 1980).
13. Lord F.M. *Applications of Item Response Theory to Practical Testing Problems* (Erlbaum, Mahwah, NJ, 1980).
14. Hambleton R.K., Swaminathan H. & Rogers H.J. *Fundamentals of Item Response Theory* (Sage Publications, Thousand Oaks, CA, 1991).
15. Ashford J.W. & Schmitt F.A. Modeling the time course of Alzheimer dementia. *Curr. Psychiatry Rep.* **3**, 20–28 (2001).
16. Novakovic A.M., Krekels E.H.J., Munafo A., Ueckert S. & Karlsson M.O. Application of item response theory to modeling of Expanded Disability Status Scale in Multiple Sclerosis. *Am. Assoc. Pharm. Sci. J.* **19**, 172–179 (2017).
17. Monsch A.U., Stähelin H.B., Spiegel R., Miserez A.R., Gass A. & Regeniter A. *Study Protocol for the BASEL Project* (Basel Study on the Elderly, Basel, Switzerland, 2007).
18. Schmid N.S., Taylor K.I., Foldi N.S., Berres M. & Monsch A.U. Neuropsychological signs of Alzheimer's disease 8 years prior to diagnosis. *J. Alzheimers Dis.* **34**, 537–546 (2013).
19. Morris J.C., *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). *Neurology* **39**, 1159–1165 (1989).
20. Benton A.L. & Hamsher K. *Multilingual Aphasia Examination 2nd ed* (AJA Associates, Iowa City, IA, 1989).
21. Reitan R. Validity of the Trail Making Test as an indicator of organic brain damage. *Percept. Mot. Skills* **8**, 271–276 (1958).
22. Delis D.C., Kramer J.H., Kaplan E. & Ober B.A. *California Verbal Learning Test: Adult Version* (Psychological Corporation, San Antonio, TX, 1987).
23. Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects (World Medical Association, Helsinki, Finland, 1964, amended 1975–2013).
24. Strittmatter W.J., *et al.* Apolipoprotein E: High-avidity binding to β-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci. USA* **90**, 1977–1981 (1993).
25. Roses A.D. Apolipoprotein E alleles as risk factors in Alzheimer's disease. *Annu. Rev. Med.* **47**, 387–400 (1996).
26. Samejima F. Graded response model. In *Handbook of Modern Item Response Theory* (Springer, New York, NY, 1997).
27. Curtis S.M.K. BUGS Code for Item Response Theory. *J. Stat. Softw.* **36**, 1–34 (2010).
28. Lord F.M. & Novick M.R. *Statistical Theories of Mental Test Scores* (Addison Wesley, Reading, MA, 1968).
29. Bornkamp B. Functional uniform priors for nonlinear modeling. *Biometrics* **68**, 893–901 (2012).
30. Lunn D., Thomas A., Best N. & Spiegelhalter D. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337 (2000).
31. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria (2003).
32. Carpenter B., *et al.* Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **76**, 1–32 (2017).
33. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org> (2008).
34. Gelman A., Rubin D.B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**, 457–472 (1992).
35. Robert C.P. & Casella G. *Monte Carlo Statistical Methods* (Springer, New York, NY, 2004).
36. Farrer L.A., *et al.* Effects of age, sex, and ethnicity on the association between Apolipoprotein E genotype and Alzheimer disease. *JAMA* **278**, 1349–1356 (1997).
37. Beck I.R., Gagneux-Zurbriggen A., Berres M., Taylor K.I. & Monsch A.U. Comparison of verbal episodic memory measures. Consortium to Establish a Registry for Alzheimer's Disease – Neuropsychological Assessment Battery (CERAD-NAB) versus California Verbal Learning Test (CVLT). *Arch. Clin. Neuropsychol.* **27**, 510–519 (2012).
38. Yang E., *et al.* Quantifying the pathophysiological timeline of Alzheimer's disease. *J. Alzheimers Dis.* **26**, 745–753 (2011).
39. Laffont C., Vandemeulebroecke M. & Concordet D. Multivariate analysis of longitudinal ordinal data with mixed effects models, with application to clinical outcomes in osteoarthritis. *J. Am. Stat. Assoc.* **109**, 955–966 (2014).

Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (http://psp-journal.com)