Computational phenotyping of cognitive decline with retest learning

Zita Oravecz[a,b,*], Joachim Vandekerckhove[c,d,e], Jonathan G. Hakun[b,f], Sharon H. Kim[a,b], Mindy J. Katz[g], Cuiling Wang[h], Richard B. Lipton[h], Carol A. Derby[h], Nelson A. Roque[a,b], and Martin J. Sliwinski[a,b]

Author final version of:

Oravecz, Z., Vandekerckhove, J., Hakun, J., Kim, S., Katz, M., Wang, C., Lipton, R. B., Derby, C. A., Roque, N. A., & Sliwinski, M. (2025). Computational phenotyping of cognitive decline with retest learning. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences.*

OBJECTIVES: Cognitive change is a complex phenomenon encompassing both retest-related performance gains and potential cognitive decline. Disentangling these dynamics is necessary for effective tracking of subtle cognitive change and risk factors for ADRD. METHOD: We applied a computational cognitive model of learning and forgetting to data from Einstein Aging Study (n = 316). EAS participants completed multiple bursts of ultra-brief, high-frequency cognitive assessments on smartphones. Analyzing response time data from a measure of visual short-term working memory, the Color Shapes task, and from a measure of processing speed, the Symbol Search task, we extracted several key cognitive markers: short-term intraindividual variability in performance, within-burst retest learning and asymptotic (peak) performance, across-burst change in asymptote and forgetting of retest gains. RESULTS: Asymptotic performance was related to both MCI and age, and there was evidence of asymptotic slowing over time. Long-term forgetting, learning rate, and within-person variability uniquely signified MCI, irrespective of age. DISCUSSION: Computational cognitive markers hold promise as sensitive and specific indicators of preclinical cognitive change, aiding risk identification and targeted interventions.

Retest learning | Computational modeling | Subtle cognitive decline | Cognitive psychometrics

Mounting evidence suggests that neuropathological changes associated with Alzheimer's Disease and Alzheimer's Disease Related Dementias (AD/ADRD) are detectable up to three decades prior to the clinical diagnosis of dementia (Sperling et al., 2011; Hadjichrysanthou et al., 2020; Verlinden et al., 2015; Ritchie et al., 2015). During this preclinical stage, cognitive and behavioral changes are subtle not only in magnitude but also in terms of the underlying cognitive processes they reflect. Assessing these subtle changes accurately in longitudinal studies is hindered by both within-person variability in performance (MacDonald et al., 2009) and retest-related effects (Wilson et al., 2006). Retest (or practice) effects refer to the ubiquitous finding that performance on cognitive tests improves with repeated testing.

It is widely recognized that retest effects can bias longitudinal estimates and intervention effects. Retest effect sizes range between 0.1 and 0.4 SD units (Calamia et al., 2012; Goldberg et al., 2015; Zelazo et al., 2014; Morrison et al., 2015; Salthouse, 2009, 2010). This retest learning can obscure years of memory decline in preclinical AD (Hall et al., 2000) and overwhelm normative cognitive aging effects, which are in the range of 0.01-0.03 SD units per year (Lipnicki et al., 2017). These effects confound the detection of cognitive change by biasing estimates of the underlying performance on a given assessment. There is currently no consensus on best practice to address retest effects.

Importantly, recent studies suggest that retest effects are not merely a source of bias but may also provide important signals for detection of subtle cognitive impairment. For example, preclinical AD is marked by a reduction in practice effects, and the magnitude of retest gain is inversely related to the risk of progressing to a clinical level of impairment (Goldberg et al., 2015; Hassenstab et al., 2015; Young et al., 2023). These findings suggest that assessment of practice effects may provide face-valid indicators of preclinical AD, with patterns of diminishing practice providing valuable insights into the early stages of cognitive impairment.

These findings underscore the potential value in characterizing cognitive retest effects and distinguishing them from long-term changes in cognitive functioning. Traditional longitudinal studies obtain cognitive assessments at widely-spaced intervals (e.g., annually), making it challenging or impossible to cleanly separate retest effects from long-term cognitive change (Hoffman et al., 2011; Rentz et al., 2013; Mortamais et al., 2017). Strategies like the measurement burst design (Sliwinski, 2008) can play a crucial role in disentangling these effects and improving the reliability of studies aimed at detecting and understanding subtle cognitive changes. Measurement bursts involve the administration of ultra-brief cognitive tests at high frequency – often multiple times within a short period (e.g., weekly or monthly). By conducting assessments in relatively quick succession,

[a] Department of Human Development and Family Studies, Pennsylvania State University; [b] Institute for Computational and Data Sciences, Pennsylvania State University; [c] Department of Cognitive Sciences, University of California, Irvine; [d] Department of Statistics, University of California, Irvine; [e] Department of Logic and Philosophy of Science, University of California, Irvine; [f] Department of Neurology, Pennsylvania State University; [g] Department of Neurology, Albert Einstein College of Medicine; [h] Department of Epidemiology and Population Health, Albert Einstein College of Medicine

[*] Correspondence concerning this article should be addressed to Zita Oravecz (zita@psu.edu).

This work was supported by National Science Foundation grant #2051186.

measurement bursts capture short-term fluctuations and trends in cognitive performance and allow us to distinguish immediate retest effects from more lasting changes in cognitive abilities. This improves the granularity and ecological validity with which learning and forgetting processes can be captured as they unfold over time (Moore et al., 2017; Singh et al., 2023). Moreover, this design also captures day-to-day variability in cognitive performance, which has itself been linked to unhealthy aging (Cerino et al., 2021).

In the present study we perform model-based computational phenotyping (Patzelt et al., 2018) to quantify individual differences in cognitive retest effects and isolate them from longer-term changes in cognitive ability. We make use of a multi-timescale learning process model to decompose cognitive performance time-series data—collected across multiple bursts of multiple days of high-frequency assessment—into a set of theoretically meaningful and behaviorally interpretable model parameters, or cognitive markers; for example, change in asymptotic (peak) performance from burst to burst (cognitive decline), performance inconsistency (variability), rate of retest improvement (learning), and loss of retest advantages between measurement waves (forgetting). These indicators go beyond summarizing data in simple statistical measures, and the approach has the potential to reveal new markers of subtle cognitive change and risk factors for ADRD.

Since the goal of proposing this methodological tool is the early identification of ADRD risk, we selected tasks that assess cognitive domains, such as processing speed (e.g., Symbol Search task), and working memory (e.g., Color Shapes task) that are implicated in preclinical AD specifically. Notably, meta-analytic results demonstrate that working memory and processing speed have equivalent rates of decline to episodic memory among cognitively normal individuals with elevated beta-amyloid (Baker et al., 2016). Previous work also showed that ambulatory measures of associative memory, processing speed, and working memory were associated with AD biomarker levels at baseline to a similar degree as conventional cognitive measures (Han et al., 2017; Nicosia et al., 2023). Although our methodology does not directly measure episodic memory, it incorporates statistical models that examine rates of learning and forgetting, which are core features of early cognitive decline in AD/ADRD. By leveraging intensive longitudinal data, these models provide insight into cognitive processes that traditional measures of episodic memory typically assess, thereby addressing critical aspects of early-stage decline.

The core of our approach to analyzing high-frequency cognitive assessments is a computational process model of learning, the Bayesian double exponential model (BDEM). In previous studies, double exponential learning models have shown good fit to practice effects in repeated cognitive testing (Broitman et al., 2019; Munoz et al., 2015). There is also evidence that the multilevel generalization of this approach can reveal individual differences in learning processes that are associated with age and mild cognitive impairment (Oravecz et al., 2022). Using this approach asymptotic performance is an optimized measure of best possible performance, and change from year-to-year in asymptotic performance represents a practice effect free assessment of within-person change over time. This "change in asymptote" parameter captures across-bursts differences relative to the person’s initial asymptote, so that the current model already controls for the participant’s baseline task ability through a model parameter. In the present study, we extend this modeling approach with an account of the degree to which individuals experience loss of retest gains across measurement bursts. We interpret this new parameter as reflecting the rate of long-term ‘forgetting’ of retest related gains that occurs between measurement bursts (i.e., over a one-year interval). The current study was designed to examine whether retest learning features capture relevant phenotypic information (i.e., memory impairment) that is common among adults with age-associated mild cognitive impairments (MCI).

Methods

Participants and sampling procedures. The Einstein Aging Study (EAS) is an ongoing longitudinal study of risk factors for MCI and dementia. Participants for the EAS were recruited via registered voting lists in Bronx County, NY. All are English- speaking, community-residing, ambulatory, and aged 70 years and over. All participants provided written informed consent, and the study was approved by the Albert Einstein College of Medicine Institutional Review Board. In the current analysis, we had 316 participants, of whom 86 (27.2%) completed one burst, 31 (9.8%) completed two bursts, 51 (16.1%) completed three bursts, 76 (24.1%) completed four bursts, 50 (15.5%) completed five bursts, and 22 (7.0%) completed six bursts. This study engages in ongoing recruiting, so the variable number of bursts reflects, in part, time since enrollment. The mean age (standard deviation in parentheses) of the sample at baseline was 77.54 (4.98) years and 67% were female ($n = 105$ male, and $n = 211$ female). The sample was racially and ethnically diverse with 46.2% ($n = 146$) identifying as non-Hispanic Whites, 39.9% ($n = 126$) as non-Hispanic Blacks, 9.8% ($n = 31$) as Hispanic Whites, 2.9% ($n = 9$) as Hispanic Blacks, 1.0% ($n = 3$) as Asian, and 0.3% ($n = 1$) as more than one race/ethnicity. The mean education of the sample was 15.09 (3.55) years. Based on the neuropsychological assessment and Jak-Bondi criteria (Jak et al., 2009), 29.1% ($n = 92$) of participants were classified as having MCI at baseline. In the analysis we included (a) age at Burst 1, (b) MCI status at Burst 1, (c) sex, (d) number of years of education, (e) race, and (f) ethnicity.

The EAS follows a measurement burst design consisting of repeated bursts of ambulatory ecological momentary assessments and clinic-based neuropsychological evaluations and collection of demographic information. During an ambulatory burst of 16 days, participants completed six brief ses-

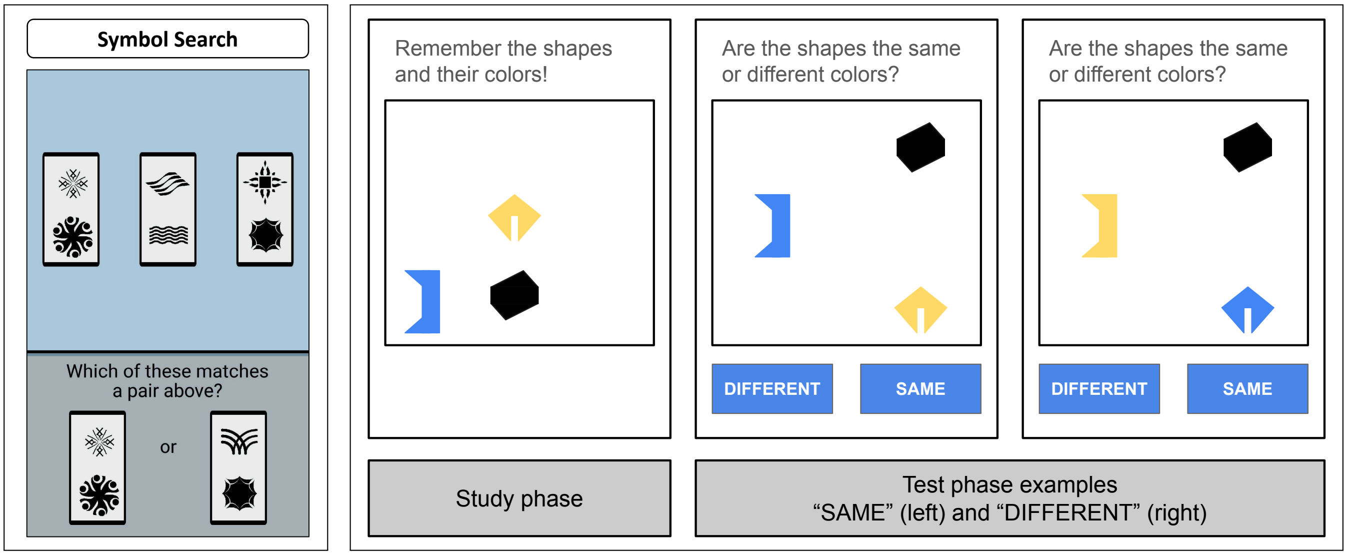
Symbol Search
Which of these matches a pair above?
or
Remember the shapes and their colors!
Are the shapes the same or different colors?
DIFFERENT
SAME
Are the shapes the same or different colors?
DIFFERENT
SAME
Study phase
Test phase examples
“SAME” (left) and “DIFFERENT” (right)

Fig. 1. Illustrations of the two cognitive tasks in the Einstein Aging Study. **Left:** A trial from the Symbol Search task. **Right:** A trial from the Color Shapes task.

sions (up to five minutes each) per day on a study-provided smartphone, during their typical waking hours in daily life settings. These brief sessions were composed of cognitive assessments (‘brain games’) and brief self-reports about their daily experiences (not analyzed in the current study). The middle four of the six sessions were prompted by beeps and were scheduled approximately 3.5 hours apart, with times varying randomly across the days of the week. Morning and evening sessions were self-responding. After each burst, participants returned the study smartphone at a clinic visit for data download. Although the study aimed for annual follow-ups, the number of years passed between bursts varied across bursts and people ($M = 1.01$, $SD = 0.30$). We note that our modeling approach took into account the between and within-person variations in elapsed time between bursts, that is, it accounts for irregularly-spaced bursts.

Materials. While several cognitive domains are measured in the EAS, here we focus on RT data from the Symbol Search task, measuring processing speed. We analyzed daily aggregates of RTs. To provide some evidence on the robustness of these findings, we also briefly summarize results from analyzing RT data from the Color Shapes task (measuring visual working memory) with the same model.

Demographics. The demographic measures were coded based on self-reports from the participants via a questionnaire. In the current analyses we used age (in years, standardized for the analysis), sex (male/female, male as reference), education (years in school, standardized), and ethnicity (Caucasian/African American/Hispanic White/Hispanic Black/Asian/Other).

Mild cognitive impairment status. All participants underwent neuropsychological assessment to determine their cognitive status, including measures of memory, executive function, attention, language, and visuospatial ability. MCI status was classified by the Jak-Bondi criteria (Jak et al., 2009). In short, to be classified as MCI, the participant needed to have impaired scores on at least two measures of the same cognitive domain or impaired score in at least three out of five cognitive domains; or they needed to show functional decline, as assessed by the Instrumental Activities of Daily Living Scale (Lawton & Brody, 1969), where impairment was defined as scores of one standard deviation below the sex-, age- and education- adjusted normative mean.

The Symbol Search task. The Symbol Search task, shown on the left side of Figure 1 measures processing speed. In the current study, on each trial of the task, participants saw three symbol pairs at the top of the screen and two symbol pairs at the bottom of the screen. They were instructed to match as quickly and accurately as they could one of the two pairs presented at the bottom to one of the three pairs at the top. Participants completed 11 trials per session. We analyzed daily aggregates of correct-trial RTs with the BDEM.

The Color Shapes task. The Color Shapes task illustrated on the right side of Figure 1 captures visual short-term working memory binding task and has been shown to be sensitive to cognitive status and early risk for ADRD in cross- sectional studies. Participants are asked to memorize the color and shape of objects and then judge whether a subsequent probe is ‘same’ or ‘different’. Participants completed seven trials per session. We analyzed daily aggregates of RTs with the BDEM.

Statistical Analysis. In order to illustrate our analytical approach, Figure 2 shows a graphical representation of response times (RTs) over measurement occasions from a synthetic individual participating in three measurement

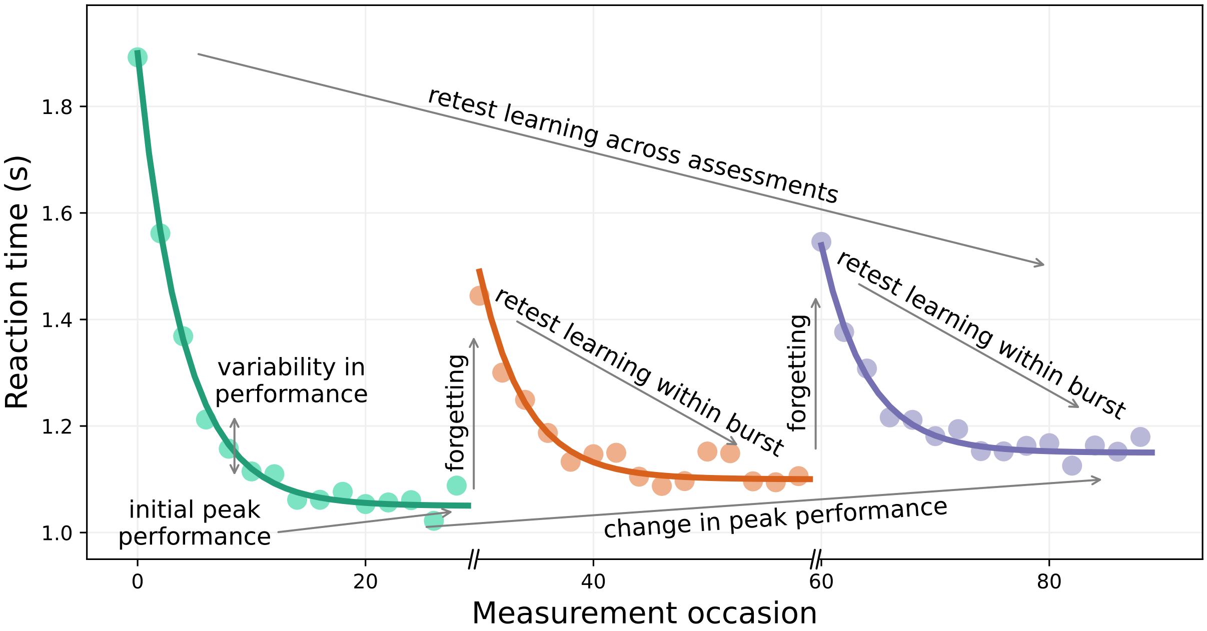

Fig. 2. Illustration of the Bayesian double exponential model with three bursts of Symbol Search data for one synthetic participant. Data are displayed as dots colored green for Burst 1, orange for Burst 2 and purple for Burst 3. Corresponding negative exponential curves (solid lines) show model fit. The model captures multiple trends at once, including within-burst retest learning (each exponential curve declines), forgetting (each new curve starts higher than the previous one ended), across-burst retest learning (worse performance in the beginning of the study than at the end), and across-burst change in peak performance (the asymptote of each exponential curve is higher than the last).

bursts. Bursts are depicted in different colors, and each dot in the figure refers to the average RT of all trials (in all sessions) on a given day.

Figure 2 shows multiple competing processes generating the displayed RT data based on the Bayesian double negative exponential modeling approach. Globally, we see that in the beginning of the study RTs are slower (around 1.9 s) than at the end of the study (under 1.2 s) – an improvement in cognitive performance likely results from retest learning across all the assessments in the study. At the same time, within each burst, participants' initial RTs are slower than those near the end of the burst, as participants become more practiced at the task, representing retest learning within burst. To account for these simultaneous and competing processes, the BDEM allows for an overall, slow-timescale improvement in cognitive performance that takes place across the study, as well as a quicker 'warm-up' improvement within a burst that starts from the second burst on. These across-burst and within-burst improvements are modeled through two separate negative exponential functions. Importantly, long-term (slow-timescale) cognitive change can then be modeled as changes in the within-burst peak (asymptotic) performance from one burst to the next. Peak performance for reaction times corresponds to low values, and cognitive decline manifests as slowing in peak performance (i.e., increase in reaction time values).

While there is improvement (retest learning) within each burst, some of that improvement is subsequently lost between bursts. This is shown by the difference in performance in the beginning of each new burst compared to the peak performance of the previous burst. We term this loss of retest gain 'forgetting', as indicated by the upward arrows in Figure 2. The model also accounts for performance inconsistency by capturing within-person variability in performance.

In our current application, we examine the association between individual differences in cognitive features extracted with the BDEM and cognitive status (presence/absence of MCI at baseline), as a first step towards establishing the usefulness of this approach for detection of AD/ADRD risk. We will demonstrate how BDEM can capture six cognitive markers at once: peak performance, change in peak performance (long-term change), forgetting between bursts, learning rate across the study, learning rate within bursts, and within-person variability in performance; all using data from the EAS. Mathematical details of the BDEM approach are provided in the Appendix.

To demonstrate the usefulness of the BDEM approach, we first did a simple analysis based on difference scores to explore the data. As most people had only two bursts of data, we focused on assessing the performance changes between Burst 1 and Burst 2. We computed the average reaction times for each participant in both bursts and then determined the change in their performance by calculating a difference score (Burst 2 minus Burst 1). Our findings revealed an average improvement of 0.14 seconds in reaction times ($M = -0.14$, with a 95% CI of $[-0.18, -0.10]$). This improvement was statistically significant, as indicated by a t test ($t = -6.41$, $df = 229$, $p = 8.38e - 10$), suggesting that participants generally became faster at the task over the course of a year. Since it is unlikely that our elderly partici-

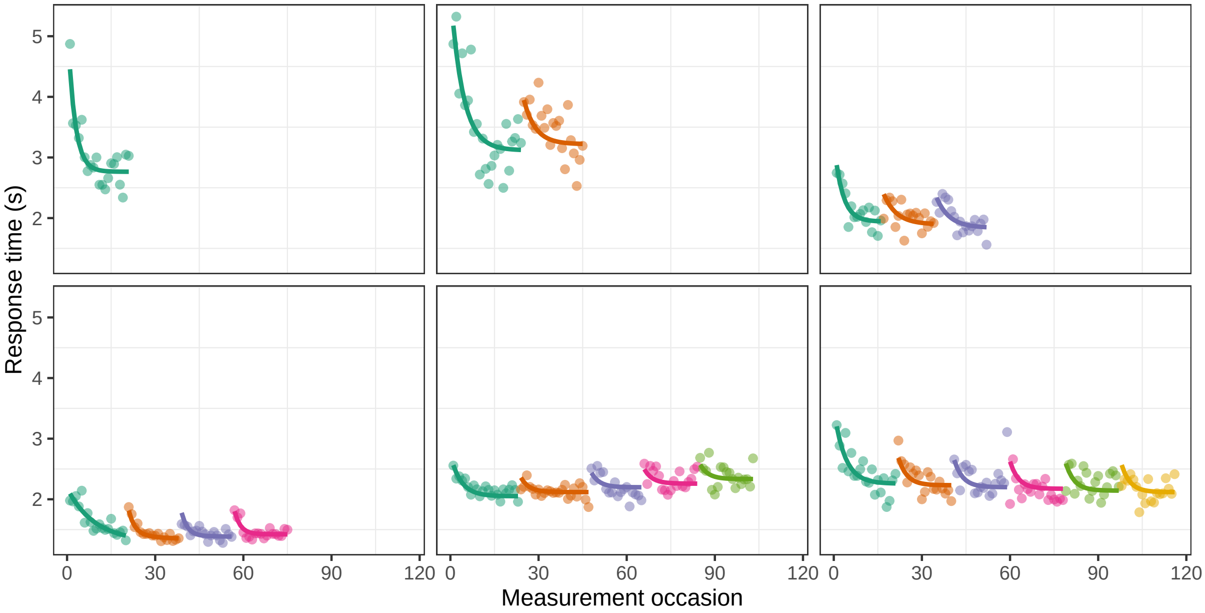

Fig. 3. Model fit to participant-level task data. Each of the six panels presents data from a different participant from the Einstein Aging Study with varying number of bursts. Dots are daily response time aggregates and model fit is shown as a solid line.

pants on average would improve in terms of their processing speed in approximately a year (which is the typical time difference between the two bursts), these results highlight the importance of accounting for retest learning effects. Later analysis with the BDEM shows that in fact there is a credible decline in processing speed across the year when retest learning effects are statistically unconfounded.

Results

Data analysis scripts are available on OSF (`osf.io/v3dnc`). To quantify model fit, we calculated the R^2 statistic, which captures the proportion of variance in RTs explained by our model. R^2 was 0.85, indicating a very good fit of the BDEM for the Symbol Search data. For a visual illustration of model fit, we plotted model-predicted trajectories over the data points, for every person. This is illustrated for six participants, with varying number of bursts, in Figure 3. As can be seen, the curves generated by the model closely resemble key aspects of individual-level data, including (1) alignment of the exponential curve's height with the observed initial data points, (2) convergence of the exponential curve's asymptote with performance near the end of each burst, and (3) a pattern of change in performance across observations that exhibits an exponential shape.

Group-level descriptions. Group-level summaries for the Symbol Search task are shown in Table 1 in terms of means and corresponding 95% credible intervals (CI). In the Bayesian framework, 95% credible intervals capture a range in which a parameter estimate falls with 95% probability, quantifying the uncertainty around the parameter estimate. For every cognitive marker, two descriptions are shown: the population mean across all participants, and the population standard deviation (SD), which captures the magnitude of individual differences in the given marker. Table 1 shows that on average the peak performance RT was 2.59 s, with considerable heterogeneity across participants ($M = 0.86$). Change in peak performance was positive on average, around 0.07 s per year ($M = 0.07$, with 95% CI [0.01, 0.13] not containing 0), representing a general decline in performance across bursts (slower RTs). However, there was inter-individual variability with regard to peak performance change as well ($M = 0.21$).

There was a considerable amount of forgetting between bursts on average ($M = 0.48$), with substantial individual differences in this feature ($M = 0.28$): some participants forgot very little between burst, whereas for others the effect of forgetting was as large as 1 s. Learning rates across the whole study and within burst were similar ($M = 0.41$ and $M = 0.49$, respectively), with some inter-individual variability in both. Finally, the performance inconsistency (within-person variability) was 0.67 s on average, also with considerable heterogeneity ($M = 0.45$). In the next subsection, we discuss and test possible exogenous sources of individual differences in our digital cognitive markers.

Associations between cognitive markers and person--level predictors of the Symbol Search task. Individual differences in these six cognitive markers might be meaningfully linked to other person-level characteristics, such as mild cognitive impairment. We tested these associations by regressing our cognitive markers on a set of predictors: age, MCI status, sex (with male as reference), education (in years), race (with White as reference) and ethnicity (with

Table 1. Group level results. Posterior means and the boundaries of 95% credible intervals of group-level mean and standard deviation (SD) of the six key cognitive markers based on data from the Symbol Search task.

Group-level cognitive marker	Mean	Quantiles	
		2.5%	97.5%
Mean peak performance	2.59	2.40	2.79
SD of peak performance	0.86	0.77	0.95
Mean change in peak performance	0.07	0.01	0.13
SD of change in peak performance	0.21	0.19	0.24
Mean forgetting between bursts	0.48	0.40	0.58
SD of forgetting between bursts	0.28	0.25	0.32
Mean learning rate across bursts	0.41	0.32	0.50
SD of learning across bursts	0.28	0.23	0.33
Mean learning rate within bursts	0.49	0.39	0.60
SD of learning within bursts	0.19	0.15	0.23
Mean within-person variability	0.67	0.57	0.78
SD of within-person variability	0.45	0.41	0.50

Note: SD = standard deviation

Table 2. Summary of selected results on associations between cognitive markers and person-level predictors for the Symbol Search task. Posterior means and the boundaries of 95% credible intervals of the regression coefficients linking the cognitive markers to the person-level predictors.

Cognitive marker	Person-level predictor	Mean	Quantiles	
			2.5%	97.5%
Peak performance	MCI status	0.82	0.60	1.06
	Age	0.12	0.01	0.22
	Education	-0.17	-0.28	-0.06
Within-person variability	MCI status	0.31	0.20	0.43
	Age	0.02	-0.04	0.07
	Education	-0.09	-0.14	-0.03
Forgetting between bursts	MCI status	0.22	0.11	0.33
	Age	-0.02	-0.06	0.02
Learning rate within burst	MCI status	-0.10	-0.19	-0.02
	Age	-0.01	-0.05	0.03
	Sex	-0.12	-0.22	-0.03

non-Hispanic as reference). These regression parameters were all embedded in a single BDEM for a one-step analysis, therefore all reported results are partial regression coefficients, meaning they summarize the effect of one conditional on all the others.

Selected results based on Symbol Search data are displayed in Table 2. The table shows all regression coefficients for which the 95% credibility interval did not contain 0 (i.e., credible associations) and some associations related to age. All results are shown in the Appendix.

Individual differences in processing speed peak performance – a cognitive marker that quantifies performance disentangled from retest effects—were credibly related to MCI status ($M = 0.82$), age ($M = 0.12$), and education ($M = -0.17$). As expected, both older participants and participants with MCI had slower peak RTs – however, the standardized effect size (normalized using the group-level SD shown in Table 1) was comparatively large for MCI status (0.82 seconds, which is $0.82/0.86 = 0.95$ in terms of standardized effect size based on the group-level peak performance SD shown in Table 1) but small for age ($0.12/4.98 = 0.02$ seconds per year or in standardized effect size $0.12/0.86 = 0.14$). Finally, participants with more years of education tended to have faster RTs.

In addition to being able to disentangle long-term cognitive change in peak performance from retest effects, with the BDEM we may also capture individual differences in learning and forgetting processes and we may attempt to explain these individual differences with exogenous variables. In our data, the amount of forgetting between bursts was larger for participants with MCI status – by around 0.22 s on average. Given the group-level SD of 0.28 ($0.22/0.28 = 0.79$), this is a large effect. As can also be seen from Table 2, the predictive link between age and forgetting was practically 0 ($M = -0.02$, with a very narrow 95% CI around it: $[-0.06, 0.02]$). It is exactly this differential predictive link—with MCI but not age—that makes this novel cognitive marker a promising candidate for an early, specific marker of cognitive impairment not due to normative aging. Similarly, the learning rate within a burst was also only related to MCI status ($M = -0.10$) and not to age ($M = -0.01$), where participants with MCI showed slower rates of learning (a medium sized effect, $0.10/0.19 = 0.53$). Sex was also related to learning rate, with female participants showing slower learning within bursts ($M = -0.12$). The BDEM captured individual differences not only in learning processes but also in performance inconsistency. Interestingly, performance inconsistency (within-person variability) was also selectively linked only to MCI status ($M = 0.31$) and not to age ($M = 0.02$). Participants with more years of education also showed less performance inconsistency ($M = -0.09$).

The difference between participants with and without MCI is further highlighted in Figure 4. As can be seen, participants with MCI (dashed line) had slower peak performance RTs (higher RT values). Looking at the distance between an asymptote (peak performance) and the top of the next exponential curve, it is also apparent that participants with MCI had larger rates of forgetting. The increased variability in performance is shown by the shading around the predicted curves. Finally, the slower learning rate for the MCI group meant that it took longer for them to reach their peak (asymptotic) performance.

A conceptual replication: Associations between cognitive markers and person-level predictors in the Color Shapes task. To demonstrate the robustness of the above findings, we replicated the analysis on RT data from the same study using another cognitive task (Color Shapes) that captured a different cognitive domain. The Color Shapes task assesses working memory, specifically short-term memory binding. Changes in working memory capacity to build, maintain, and rapidly update arbitrary bindings, such as features (e.g., color) to items (e.g., shapes), could be early markers of ADRD onset (Parra et al., 2022).

Results are shown in Table 3, with links that were not replicated typeset in italics (compared to Symbol Search

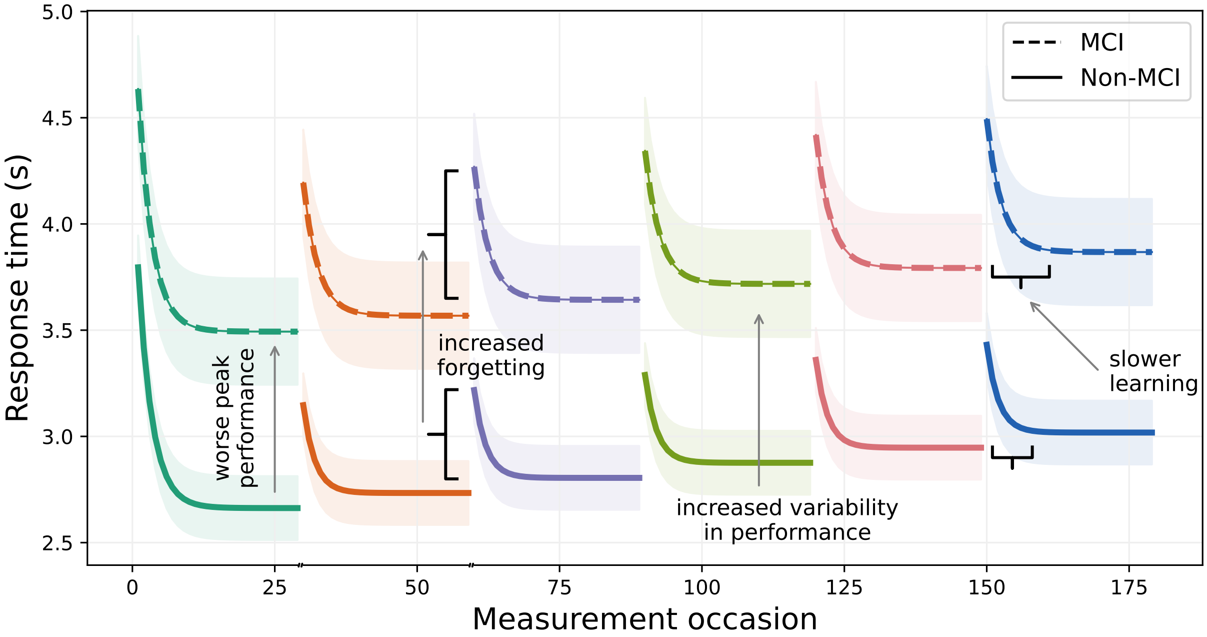

Fig. 4. Visualization of model-predicted group differences. Model-predicted trajectory estimates for participants with and without MCI, with the four key credible differences between the groups highlighted: Participants with MCI had worse peak performance, increased forgetting, increased variability in performance, and slower learning within bursts.

results shown in Table 2), with two new links in the bottom two rows. As can be seen, the associations between peak performance and age (lower with age) and MCI (lower with positive status) were replicated, but we did not find a credible association between years of education and peak performance on this task. Importantly, we replicated the key findings regarding MCI status being selectively associated with within-person variability and forgetting between bursts, while age was again not credibly linked to either of these two. Lastly, the credible links between learning rate and MCI status and sex did not replicate, but two new links emerged: participants with more years of education tended to have less forgetting between bursts and learning rate was slightly increased with age.

Table 3. Summary of selected results on associations between cognitive markers and person-level predictors for the Color Shapes task. Posterior means and the boundaries of 95% credible intervals of the regression coefficients linking the cognitive markers to the person-level predictors.

Cognitive marker	Person-level predictor	Mean	Quantiles 2.5%	97.5%
Peak performance	MCI status	0.20	0.05	0.35
	Age	0.11	0.05	0.18
	Education	*0.01*	*-0.06*	*0.08*
Within-person variability	MCI status	0.15	0.06	0.24
	Age	0.02	-0.02	0.06
	Education	-0.07	-0.11	-0.02
Forgetting between bursts	MCI status	0.17	0.07	0.28
	Age	0.02	-0.03	0.06
Learning rate within burst	*MCI status*	*0.06*	*-0.04*	*0.17*
	Age	-0.01	-0.05	0.04
	Sex	*0.00*	*-0.09*	*0.09*
Forgetting between bursts	Education	-0.05	-0.10	-0.01
Learning rate across bursts	Age	0.06	0.01	0.10

Discussion

In the present study, we showed how to dissociate retest effects from longer-term cognitive changes in high-frequency cognitive assessment designs with the BDEM. Building on previous evidence that the BDEM precisely captures patterns of retest learning and levels of peak ('best') performance achieved within a measurement burst (Oravecz et al., 2022), here we showed that this model can be extended to capture performance dynamics that occur across multiple (and possibly irregularly spaced) measurement bursts over time. This framework allowed us to extract individual-specific cognitive markers of retest learning that were meaningfully linked to person-level characteristics. Specifically, the inclusion of a forgetting parameter allowed us to capture the degree of retest gain that is lost between bursts. Interindividual differences in the forgetting parameter in older adults were selectively predictive for mild cognitive impairment, but not age, indicating that this parameter may be sensitive to cognitive phenotypic information relevant to AD/ADRD risk (i.e., amnestic patterns).

Mild cognitive impairment status was also credibly associated with three other features extracted with the BDEM: peak performance, learning rate, and within-person variability, while age was only associated with peak performance. However, neither MCI status nor age were meaningfully related to long-term changes in cognitive peak performance in the current study. This further highlights the importance of capturing the subtle latent processes of cognitive change.

We propose that parameters of the BDEM should be studied for their use as novel cognitive markers of subtle cognitive change during AD/ADRD-related pathological processes. It is furthermore important to highlight that, while we interpret these parameters as reflecting learning and memory processes (including the forgetting parameter), we derive estimates of these parameters from RT data collected during a task designed to assess processing speed/attention. Use of the BDEM in this way may open new avenues for assessing multiple domains/dimensions of cognitive health using a single-task paradigm, which would improve efficiency of cognitive assessment and reduce participant/patient burden in future trials.

The conceptual replication revealed similar key associations between these cognitive markers and MCI status in the context of a working memory task. Specifically, the same participants with MCI status also tended to have lower peak performance, higher performance inconsistency and higher levels of forgetting on the Color Shapes task as well. While these findings are based on the RT measure of this task, as opposed to the more conventionally used accuracy-based measure (Parra et al., 2010), the BDEM is not limited to RT data. In fact, previous work (Oravecz et al., 2022) showed associations between age and BDEM features based on an error distance measure of another working memory task (Dot Memory). Future work could also explore simultaneous analysis of RT and accuracy data by combining the BDEM with a hierarchical diffusion modeling approach (Vandekerckhove et al., 2011).

A limitation of our study concerning the results related to MCI status is that we focused solely on baseline, without considering possible changes in MCI status over time. MCI classification serves to characterize dementia risk but is imperfect. Indeed, some individuals classified as MCI at baseline may 'revert' to non-MCI status at follow-up assessments. This reversion could stem from various factors, including measurement error or the alleviation of transient conditions (e.g., medication side effects, dehydration or nutritional deficiencies, stress) contributing to cognitive decline. Reversion from MCI baseline status could also be driven by retest-related performance boosts that result in individuals with underlying impairment performing above the MCI threshold with repeated testing at follow-up, potentially leading to misclassification. Conversely, individuals not initially identified as MCI at baseline may later meet the classification threshold for MCI during subsequent assessments, indicating incident MCI. As the EAS is ongoing, with participants accruing several bursts of measurements over the years, we anticipate an increase in incident MCI rates, providing opportunities to examine the predictive value of our novel digital markers for incident MCI with greater statistical power.

A possible limitation of our approach is the measurement variability that stems from environmental factors of the participants' natural environments (see, e.g., Benson et al., 2023; Zhaoyang et al., 2022; Hyun et al., 2019, 2024). However, we emphasize that by choosing high-frequency ambulatory measurement designs (e.g., the measurement burst design in the current study), in which each individual performs many repeated measures over the course of two weeks, we rely on distributing testing across a wide range of contexts that should make our model implied outcomes robust to the testing context. Environmental factors in this framework are considered as stochastic sources of measurement error. Also, individual differences in performance variability are directly modeled through BDEM parameters (e.g., intra-individual variability and retest learning), and combined with the dense assessments across the different context can in fact highlight individual differences in contextual response. This approach also ensures that occasional extreme measurement occasions (e.g., due to distraction or interruptions) do not exert undue influence on the outcomes.

Computational modeling of high-frequency cognitive assessments, such as those conducted through smartphone-based testing, provides advantages over conventional testing methods by capturing retest effects and learning trends, and by estimating cognitive decline independently of practice effects. This is in contrast with traditional neuropsychological testing approaches, which typically capture only a snapshot of an individual's performance at a specific moment and elide natural fluctuations in performance that occur in everyday life. Consequently, these approaches may be insensitive to subtle cognitive changes marked by variability in performance, retest trends, or long-term forgetting (see, e.g., Mortamais et al., 2017; Rentz et al., 2013).

Previous studies (Elbin et al., 2023; Thompson et al., 2022; Wrzus & Neubauer, 2023) suggest that ambulatory high-frequency assessments showcased in our study fit well with work done in a neuropsychological or memory clinic. These tests provide sensitive and ecologically valid information that complements the metrics that clinicians would traditionally have in their clinical portfolio. However, future work should address designing clinical grade protocols and normative ranges of change for clinically informative outcomes. Recent developments of open-source data collection infrastructure (Hakun et al., 2024) make our approach feasible for such future work.

Effective planning and evaluation of preventive interventions for cognitive impairment necessitate assessments capable of detecting subtle markers of susceptibility and monitoring them over time. While comprehensive neuropsychological exams are the current gold standard for detecting cognitive impairment, they are lengthy and often require in-person administration, limiting their scalability for use as endpoints in research and clinical trials. Our study demonstrates that mobile cognitive testing combined with Bayesian modeling can yield novel cognitive markers capturing learning and performance inconsistency, offering a promising approach for long-term monitoring in clinical settings and therapeutic trials. Besides being able to disentangle retest effects from longer-term cognitive change, a major advantage of the BDEM approach is the ability to quantify directly interpretable latent cognitive features that could be translated for future use in long-term monitoring in standard-of-care settings or therapeutic targets in clinical trials. In fact, the presented approach is designed to supply clinically useful information for every individual, in terms of personalized probabilities of impairment and decline based on Bayesian posterior probability distributions of the cognitive markers. Thus, we are optimistic that learning model-based computational phenotypes can streamline therapeutic discovery.

To further validate the practical utility of these digital cognitive markers, we propose exploring correlations with neuroimaging and blood-based biomarkers of neurodegeneration. For example, investigating how rates of forgetting relate to measures of cortical integrity from MRI or connectivity from fMRI could validate these markers as indicators of brain health (e.g., neuroimaging and blood-based/plasma-based biomarkers, see Davatzikos et al., 2011; Moradi et al., 2015). Similarly, examining associations between blood and plasma biomarkers and cognitive markers derived from the BDEM could offer a comprehensive view of neural health that is linked to specific and theoretically interpretable features of cognitive function, such as learning, forgetting, variability, and peak or asymptotic performance (Beydoun et al., 2023). Overall, this approach would offer a triangulated, multi-modal perspective of brain health, delivering a comprehensive overview of neural well-being and cognitive status.

References

Baker, J. E., Lim, Y. Y., Pietrzak, R. H., Hassenstab, J., Snyder, P. J., Masters, C. L., & Maruff, P. (2016). Cognitive impairment and decline in cognitively normal older adults with high amyloid-β: A meta-analysis. *Alzheimer's & Dementia*, *6*, 108–121.

Benson, L., Fleming, A. R., & Hakun, J. G. (2023). Sometimes you just can't: within-person variation in working memory capacity moderates negative affect reactivity to stressor exposure. *Cognition and Emotion*, *37*(8), 1357–1367.

Beydoun, M. A., Noren Hooten, N., Beydoun, H. A., Weiss, J., Maldonado, A. I., Katzel, L. I., & Waldstein, S. R. (2023). Plasma neurofilament light and brain volumetric outcomes among middle-aged urban adults. *Neurobiology of Aging*, *129*, 28–40.

Broitman, A., Kahana, M., & Healey, K. (2019). Modeling retest effects in a longitudinal measurement burst study of memory. *Computational Brain & Behavior*, *3*, 1–13.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Cerino, E., Katz, M., Wang, C., Qin, J., Gao, Q., Hyun, J., & Sliwinski, M. (2021). Variability in cognitive performance on mobile devices is sensitive to mild cognitive impairment: Results from the einstein aging study. *Frontiers in Digital Health*, *3*.

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., & Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, *32*(12), 2322.e19–2322.e27.

Elbin, R., Durfee, K., Womble, M., Dollar, C., Elbich, D., & Hakun, J. (2023). Compliance rates and symptom exacerbation for the mobile neurocognitive health (mnch) project in adolescents and adults with concussion. *Medicine & Science in Sports & Exercise*, *55*, 495–495.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis, third edition*. Taylor & Francis.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*, 103–111.

Hadjichrysanthou, C., Evans, S., Bajaj, S., Siakallis, L. C., McRae-McKee, K., de Wolf, F., & Anderson, R. M. (2020). The dynamics of biomarkers across the clinical spectrum of Alzheimer's disease. *Alzheimer's Research & Therapy*, *12*.

Hakun, J. G., Elbich, D. B., Roque, N. A., Yabiku, S. T., & Sliwinski, M. (2024). Mobile monitoring of cognitive change (m2c2): High-frequency assessments and protocol reporting guidelines. *PsyArXiv*.

Hall, C. B., Lipton, R. B., Sliwinski, M., & Stewart, W. F. (2000). A change-point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Statistics in Medicine*, *19*, 1555–1566.

Han, S. D., Nguyen, C. P., Stricker, N. H., & Nation, D. A. (2017). Detectable neuropsychological differences in early preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology Review*, *27*(4), 305–325.

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., & Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology*, *29*, 940–948.

Helm, J. L., Castro-Schilo, L., & Oravecz, Z. (2016). Bayesian versus maximum likelihood estimation of multitrait–multimethod confirmatory factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 17–30.

Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychology and Aging*, *26*, 778–791.

Hyun, J., Lovasi, G. S., Katz, M. J., Derby, C. A., Lipton, R. B., & Sliwinski, M. J. (2024). Perceived but not objective measures of neighborhood safety and food environments are associated with longitudinal changes in processing speed among urban older adults. *BMC Geriatrics*, *24*(1), 551.

Hyun, J., Sliwinski, M. J., & Smyth, J. M. (2019). Waking up on the wrong side of the bed: The effects of stress anticipation on working memory in daily life. *The Journals of Gerontology: Series B*, *74*(1), 38–46.

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., & Delis, D. C. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *The American Journal of Geriatric Psychiatry*, *17*, 368–375.

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist*, *9*(3.1), 179–186.

Lipnicki, D. M., Crawford, J. R., Dutta, R., Thalamuthu, A., Kochan, N. A., Andrews, G., ... others (2017). Age-related cognitive decline and associations with sex, education and apolipoprotein E genotype across ethnocultural groups and geographic regions: a collaborative cohort study. *PLoS Medicine*.

MacDonald, S. W., Li, S.-C., & Bäckman, L. (2009). Neural underpinnings of within-person variability in cognitive functioning. *Psychology and Aging*, *24*, 792–808.

Moore, R. C., Swendsen, J., & Depp, C. A. (2017). Applications for self-administered mobile cognitive assessments in clinical research: A systematic review. *International Journal of Methods in Psychiatric Research*, *26*.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811914008131 (Retrieved from)

Morrison, G. E., Simone, C. M., Ng, N. F., & Hardy, J. L. (2015). Reliability and validity of the neurocognitive performance test, a web-based neuropsychological assessment. *Frontiers in Psychology*, *6*.

Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., ... Ritchie, K. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's & Dementia*, *13*(4), 468–492.

Munoz, E., Sliwinski, M. J., Scott, S. B., & Hofer, S. (2015). Global perceived stress predicts cognitive change among older adults. *Psychology and Aging*, *30*(3), 487.

Nicosia, J., Aschenbrenner, A. J., Balota, D. A., Sliwinski, M. J., Tahan, M., Adams, S., ... others (2023). Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. *Journal of the International Neuropsychological Society*, *29*(5), 459–471.

Oravecz, Z., Harrington, K. D., Hakun, J. G., Katz, M. J., Wang, C., Zhaoyang, R., & Sliwinski, M. J. (2022). Accounting for retest effects in cognitive testing with the bayesian double exponential model via intensive measurement burst designs. *Frontiers in Aging Neuroscience*, 128.

Parra, M. A., Abrahams, S., Logie, R. H., Méndez, L. G., Lopera, F., & Della Sala, S. (2010). Visual short-term memory binding deficits in familial Alzheimer's disease. *Brain*, *133*(9), 2702–2713.

Parra, M. A., Calia, C., Pattan, V., & Della Sala, S. (2022). Memory markers in the continuum of the Alzheimer's clinical syndrome. *Alzheimer's Research & Therapy*, *14*(1), 1–16.

Patzelt, E. H., Hartley, C. A., & Gershman, S. J. (2018). Computational phenotyping: Using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience*, *1*.

Rentz, D. M., Parra Rodriguez, M. A., Amariglio, R., Stern, Y., Sperling, R., & Ferris, S. (2013). Promising developments in neuropsychological approaches for the detection of preclinical Alzheimer's disease: A selective review. *Alzheimer's Research & Therapy*, *5*, 58.

Ritchie, K., Ritchie, C. W., Jaffe, K., Skoog, I., & Scarmeas, N. (2015). Is late-onset Alzheimer's disease really a disease of midlife? *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *1*(2), 122–130.

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*, 507–514.

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*, *24*, 563–572.

Singh, S., Strong, R., Xu, I., Fonseca, L., Hawks, Z., Grinspoon, E., & Germine, L. (2023). Ecological momentary assessment of cognition in clinical and community samples: Reliability and validity study. *Journal of Medical Internet Research*, *25*.

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass*, *2*.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., … others (2011). 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*, *7*(3), 280–292.

Stan Development Team. (2023). *RStan: the R interface to Stan.* (R package version 2.26.22)

Thompson, L., Harrington, K., Roque, N., Strenger, J., Correia, S., Jones, R., … Sliwinski, M. (2022). A highly feasible, reliable, and fully remote protocol for mobile app-based cognitive assessment in cognitively healthy older adults. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *14*.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62.

Verlinden, V. J., van der Geest, J. N., de Bruijn, R. F., Hofman, A., Koudstaal, P. J., & Ikram, M. A. (2015). Trajectories of decline in cognition and daily functioning in preclinical dementia. *Alzheimer's & Dementia*, *12*.

Wilson, R. S., Li, Y., Bienias, J. L., & Bennett, D. A. (2006). Cognitive decline in old age: Separating retest effects from the effects of growing older. *Psychology and Aging*, *21*, 774–789.

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846.

Young, C. B., Mormino, E. C., Poston, K. L., Johnson, K. A., Rentz, D. M., Sperling, R. A., & Papp, K. V. (2023). Computerized cognitive practice effects in relation to amyloid and tau in preclinical Alzheimer's disease: Results from a multi-site cohort. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *15*.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., … Weintraub, S. (2014). NIH toolbox Cognition Battery (CB): validation of executive function measures in adults. *Journal of the International Neuropsychological Society*, *20*(6), 620–629.

Zhaoyang, R., Harrington, K. D., Scott, S. B., Graham-Engeland, J. E., & Sliwinski, M. J. (2022). Daily social interactions and momentary loneliness: The role of trait loneliness and neuroticism. *The Journals of Gerontology: Series B*, *77*(10), 1791–1802.

Appendix

Mathematical formulation of the Bayesian double exponential model. Mathematically, the Bayesian double exponential model is specified as:

$$RT_{ti} = a_i + \Delta_i B_{ti} + g_i e^{-r_i M_{ti}} + I_{(Bn_{ti}>1)} g_i^* e^{-r_i^* T_{ti}} + e_{ti}, \quad [1]$$

where RT_{ti} stands for person i's RT at measurement occasion t. The latent processes generating these RTs are modeled through key parameters that have meaningful substantive interpretations. Specifically, a_i denotes person i's asymptotic or peak performance, and Δ_i captures the person-specific linear change in this peak performance across bursts, with B_{ti} quantifying the burst start time (in years). This change in peak performance is disentangled from learning processes via the two exponential function in Equation 1. The first exponential function captures the learning process across the whole study, with M_{ti} denoting measurement time nested in study, r_i capturing the person-specific learning rate across study and g_i capturing the person-specific gain across study. The learning rate parameter quantifies the slope of the exponential, while the gain parameter captures its height. From the second burst on, denoted as $I_{(Bn_{ti}>1)}$, where Bn_{ti} is the burst number for person i at measurement occasion t, we also allow for a within-burst 'warm-up' learning process to take place, parameterized through person-specific learning rate in burst, r_i^*, and person-specific gain in burst g_i^*, with T_{ti} capturing the measurement time nested in burst. Note, however, in our Bayesian implementation of the exponential model we do not consider the within-burst gain per se, but introduce a reparameterization and instead capture 'forgetting', ϕ_i for each person i, as a new parameter. It is derived as $\phi_i = \Delta_i + g_i^* e^{-r_i^*}$. Finally, we allow for performance inconsistency via an error term e_{ti} around the exponential curve that follows a zero centered normal distribution: $e_{ti} \sim N(0, \sigma_i)$, with person-specific parameter standard deviation σ_i quantifying the amount of within-person variability.

The multilevel Bayesian implementation of the double negative exponential model. All six cognitive features in Equation 1 were allowed to differ between participants. Then the model was cast in a multilevel framework for increased estimation accuracy of the parameters and principled testing of group-level trends. This meant that all person-specific parameters (representing the cognitive features) were pooled together via population (group-level) distributions, with means regressed on predictors. For example, the person-specific peak performances a_i were assumed to follow a group-level distribution that was defined as $a_i \sim N(\mathbf{x}_i \boldsymbol{\beta}_a, \sigma_a)$, with the mean of the distribution decomposed into a product of person-specific predictor variables $\mathbf{x}_i$ and corresponding regression coefficients $\boldsymbol{\beta}_a$, with σ_a the group-level standard deviation. All other features followed the same formulation.

This model was implemented in a Bayesian statistical framework (Gelman et al., 2013). The Bayesian framework offers estimation techniques that work well for complex, nonlinear, multilevel models, in terms of higher convergence rates and admissible solutions for parameter estimates (e.g., no negative variance estimates, see Helm et al., 2016). Moreover, our multilevel Bayesian approach (Gelman & Hill, 2007) allows for simultaneous (one-step) estimation of cognitive features and regression coefficients linking them to predictors.

In the Bayesian framework every model parameter must have a prior distribution, which is combined with a likelihood function (based on the selected cognitive process model and using Bayes' rule) to yield the parameter's posterior probability distribution. Weakly informative priors were set on hyperparameters, which did not bias the estimation but focused the estimation range to plausible values (i.e., by assigning low probabilities to unreasonably high values). For example, the regression coefficient linking peak performance to age was set to a normally distributed prior with mean 0 and standard deviation 10, $\beta_{a,\text{age}} \sim N(0, 10)$. Since age entered the model as a standardized predictor, we would not expect this regression coefficient to take values above 10 in either direction (with its most likely range being between -4 and +4), and our prior reflects this knowledge.

The Bayesian model fitting for the chosen exponential model was done in Stan, which is a freely available "state-of-the-art platform for statistical modeling and high-performance statistical computation" (Carpenter et al., 2017), called from R via `rstan` (Stan Development Team, 2023). We ran four chains with 6,000 iterations each, discarding 1,000 of each as warm-up, to obtain a posterior sample size of 20,000 for every model parameter. Convergence was checked by calculating the potential scale reduction factor (Gelman et al., 2013) $\hat{R}$ and visual inspection of trace plots – no convergence problems were found. We also checked the quality of the samples via calculating effective sample sizes (number of independent pieces of information in the posterior sample) which was sufficient for every parameter (over 1,000 for 98.5% of the parameters and never under 60).

Additional results. Table 4 shows full results on associations between cognitive markers and person-level predictors for the Symbol Search task.

Table 4. Posterior means and the boundaries of 95% credible intervals of the regression coefficients linking the cognitive markers to the person-level predictors.

Cognitive marker	Person-level predictor	Mean	Quantiles	
			2.5%	97.5%
Forgetting	Intercept	0.48	0.40	0.58
	Age	-0.02	-0.06	0.02
	MCI status	0.22	0.11	0.33
	Sex	-0.02	-0.12	0.07
	Years of Education	0.00	-0.04	0.05
	Race Black	-0.01	-0.11	0.09
	Ethnic Hispanic	-0.04	-0.19	0.11
Learning rate within bursts	Intercept	0.49	0.39	0.60
	Age	-0.01	-0.05	0.03
	MCI status	-0.10	-0.19	-0.02
	Sex	-0.12	-0.22	-0.03
	Years of Education	-0.00	-0.05	0.04
	Race Black	0.05	-0.04	0.14
	Ethnic Hispanic	-0.04	-0.17	0.10
Learning rate across study	Intercept	0.41	0.32	0.50
	Age	0.00	-0.04	0.05
	MCI status	-0.01	-0.11	0.09
	Sex	0.05	-0.04	0.14
	Years of Education	-0.02	-0.06	0.03
	Race Black	-0.02	-0.11	0.08
	Ethnic Hispanic	0.01	-0.13	0.15
Peak performance	Intercept	2.59	2.40	2.79
	Age	0.12	0.01	0.22
	MCI status	0.82	0.60	1.06
	Sex	0.00	-0.21	0.22
	Years of Education	-0.17	-0.28	-0.06
	Race Black	0.09	-0.13	0.32
	Ethnic Hispanic	-0.01	-0.33	0.31
Change in peak performance	Intercept	0.07	0.01	0.13
	Age	0.01	-0.02	0.04
	MCI status	0.00	-0.07	0.08
	Sex	-0.06	-0.13	0.01
	Years of Education	-0.01	-0.04	0.02
	Race Black	-0.03	-0.10	0.03
	Ethnic Hispanic	0.02	-0.09	0.12
Within-person variability	Intercept	0.67	0.57	0.78
	Age	0.02	-0.04	0.07
	MCI status	0.31	0.20	0.43
	Sex	-0.02	-0.13	0.09
	Years of Education	-0.09	-0.14	-0.03
	Race Black	0.07	-0.05	0.19
	Ethnic Hispanic	-0.08	-0.25	0.08