Partially observable predictor models for identifying cognitive markers

Zita Oravecz[a,*], Martin Sliwinski[a], Sharon H. Kim[a], Lindy Williams[a], Mindy J. Katz[b], and Joachim Vandekerckhove[c,*]

Draft of March 6, 2025.

Repeated assessments of cognitive performance yield rich data from which we can extract markers of cognitive performance. Computational cognitive process models are often fit to repeated cognitive assessments to quantify individual differences in terms of substantively meaningful cognitive markers and link them to other person-level variables. Most studies stop at this point, and do not test whether these cognitive markers have utility for predicting some meaningful outcomes. Here, we demonstrate a *partially observable predictor* modeling approach that can fill this gap. Using this approach, we can simultaneously extract cognitive markers from repeated assessment data and use these together with demographic covariates for predictive modeling of a clinically interesting outcome in a Bayesian multilevel modeling framework. We describe this approach by constructing a predictive process model in which features of learning are combined with demographic variables to predict mild cognitive impairment, and demonstrate it using data from the Einstein Aging Study.

Partially observable predictors | cognitive psychometrics | Bayesian models

Digital technology has enabled us to collect large volumes of cognitive performance data from an individual with relative ease. For example, the well-known Project Implicit data set now contains data obtained from 2.7 million individuals for one of the dozens of cognitive performance tests (Stier, Sajjadi, Karimi, Bettencourt, & Berman, 2024). In other instances, people have shown great willingness to play 'brain games' on smartphones in daily life settings – that is, to complete brief cognitive assessments in their natural environment repeatedly during the day, for several days (Thompson et al., 2022). Such high-frequency performance data are generated by multiple underlying processes related to learning and variability in cognitive performance on various timescales (e.g., day-to-day, week-to-week). Computational cognitive psychometric modeling (Batchelder, 2010) is needed to disentangle these latent processes, and explore individual differences therein.

Over the past decades, numerous computational cognitive process models that capture the latent processes underlying observed scores of cognitive tasks have been developed. These models define cognitive parameters—unobservable or 'latent' underlying features of behavior—that can be inferred from behavioral data to understand participant performance and explain variability within and between participants. Among the more popular models are the family of drift diffusion models (Ratcliff & McKoon, 2008; Vandekerckhove, Tuerlinckx, & Lee, 2011) that separate processing speed from metacognitive factors in reaction time tasks; the expectancy-valence model (Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010) that captures risk attitudes in decision making; Rescorla-Wagner models (Browning, Behrens, Jocham, O'Reilly, & Bishop, 2015) that quantify processes in Pavlovian learning; multinomial processing tree models that quantify abilities and biases in behaviors leading to discrete outcomes (Erdfelder et al., 2009); and retest learning models that disentangle multi-timescale processes in repeated testing (Oravecz et al., 2022). The common element among these models is that they propose concrete data-generating mechanisms underlying the observed behaviors during a cognitive task. As quantitative models, they use latent variables (i.e., cognitive parameters or markers) to capture the most important characteristics of human decision making, learning, and memory.

Untangling the sources of individual differences in these latent cognitive features is a major focus of research in cognitive science. Cognitive markers have been linked to person-level characteristics such as age (Thapar, Ratcliff, & McKoon, 2003), anxiety (Charpentier, Aylward, Roiser, & Robinson, 2016), sex (Oravecz, Faust, & Batchelder, 2014), cognitive impairment (Oravecz et al., 2025), among others. However, while these studies have collectively established the validity of various cognitive markers to describe meaningful individual differences, and yielded insights that are not accessible from simple summary statistics (Yechiam, Busemeyer, Stout, & Bechara, 2005), they often do not test whether particular cognitive markers are individually predictive of meaningful criteria – for example, *clinical outcomes* such as a diagnosis of mild cognitive impairment (MCI) or Alzheimer's dementia (AD).

It is common practice to evaluate how much variance in model parameters is explained by a clinically meaningful outcome – using the clinical outcome as a predictor and parameters as the criterion (e.g., Hernaus, Gold, Waltz, & Frank, 2018; Ratcliff, Scharre, & McKoon, 2022).[1] In many interesting cases, however, it is useful to use person-specific latent cognitive markers more directly for the prediction of clinically meaningful outcomes. In this paper, our inference will be directly towards predicting whether a participant has the clinical condition, given their task behavior and performance. We will demonstrate this through the application of a

[a]The Pennsylvania State University; [b]Albert Einstein College of Medicine; [c]University of California, Irvine

All authors contributed to the final draft.

[*]Correspondence should be addressed to Zita Oravecz (zita@psu.edu) and Joachim Vandekerckhove (joachim@uci.edu).

ZO and MS were supported by the NIH (R01AG074208, R56AG074208 and U2CAG060408). MK received funding from the NIH (NIA 2 P01 AG03949), the Leonard and Sylvia Marx Foundation, and the Czap Foundation. JV was supported by the National Science Foundation (#1658303, #1850849, and #2051186).

[1] For example, we often see models of the form $\theta_p \sim N\left(\beta_0 + \beta_1 \text{MCI}_p, \sigma^2\right)$ (or nonlinear variations thereof) in which the clinical state of person p (MCI_p) is treated as known while the cognitive parameter θ_p is treated as the unknown. Realistically, both the latent parameter θ and the clinical state are unknown, and the latter must be inferred from diagnostic data.

learning model that captures practice effects in cognitive testing and show how model-based latent cognitive markers can be combined with manifest predictors into a *partially observable predictor model*. For example, rather than capturing how much variance in person-specific learning rates is explained by participants' clinical MCI status, we will formulate a partially observable predictor model for predicting the risk for MCI as an outcome.

A partially observable predictor model

Here we define the *partially observable predictor* (POP) model class. The distinguishing feature of POP models is that they use both manifest (observable) and latent (unobservable) variables in order to predict a given outcome. The latent variables are identified by participants' behavior when completing a cognitive task, and must be inferred with a generative model. The key components of the POP model are (a) a generative model for extracting latent features, and (b) a structural model to combine observable and unobservable predictors into a single predictive value. We will discuss these two component models first, before reviewing the Bayesian inference procedure we use to apply POP models to data.

Note that, throughout, we use 'manifest' and 'latent' interchangeably with 'observable' and 'unobservable,' respectively. We use 'prediction' in the statistical sense, which does not imply that the criterion happens or is observed at a later time.

Process model to identify latent features. We start by specifying a model with which we can extract latent cognitive features from high-frequency data. For repeated measures of performance scores collected from a participant i, at occasions t, we model the data y_{ti} as a function of latent cognitive markers that represent theoretically meaningful constructs. In our illustrative example, we use an exponential model of practice – a learning process model that captures practice effects in repeated cognitive testing (Heathcote, Brown, & Mewhort, 2000). However, we emphasize that this process model could take the form of any other parametric model that proposes a latent data-generating mechanisms of a set of observations – the model could be simple (e.g., a Gaussian distributions with some mean and standard deviation), but here we introduce a process model with interpretable parameters.

Our selected process model of practice effects, the exponential learning model, is specified as:

$$y_{ti} \sim N\left(a_i + g_i e^{-r_i M_{ti}}, \varepsilon_i^2\right). \qquad [1]$$

On the left hand side we have our data y_{ti}, which will be response times coming from repeated assessments with a cognitive task, from participant i, on occasions t. On the right hand side we define how these data were generated by parameterizing an assumed theoretical process. In this particular case: (1) a_i captures person i's asymptotic or peak performance, (2) g_i quantifies person i's gain in performance, (3) r_i captures person i's learning rate across measurements (with M_{ti} denoting person i's measurement time across t occasions), and finally (4) ε_i is the standard deviation of the time-and-person-specific error (noise) term, to capture person i's intra-individual variability (i.e., performance inconsistency, see, e.g., Dzierzewski et al., 2013). These four cognitive markers are illustrated graphically in Figure 1. For ease of exposition, and also to potentially abstract away the generative model itself, we will often collect all process model parameters of person i in a vector of latent variables, $\mathbf{\Lambda}_i$.

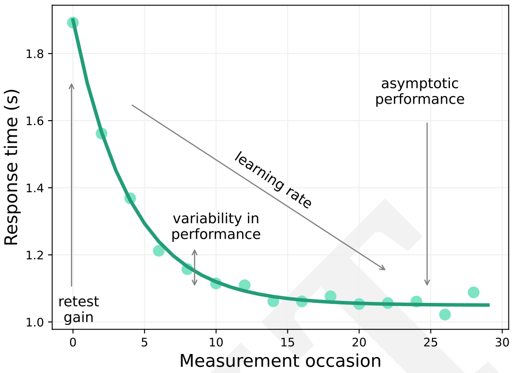

Fig. 1. The exponential learning model. The observed day-level performance data, in terms of mean response time, are represented by dots, while model fit is illustrated by a continuous negative exponential curve. This model is governed by four parameters: gain and learning rate, which control the height and the steepness/slope of the exponential curve, respectively; variability in performance, which captures the dispersion of measurements around the fitted model; and asymptotic performance, which quantifies the position of the curve's asymptote.

Structural model to combine observable and unobservable predictors. A typical next step would be to regress the latent cognitive parameters on a set of covariates or predictors. For example, considering the asymptote parameter a_i, which represents a person i's peak performance (disentangled from practice effects), researchers might want to know if individual differences in asymptotic performance on cognitive domain are meaningfully related to other person-level characteristics, such as sex, age, ethnicity, or some clinically meaningful outcome like MCI, genetically inherited AD status, or suicidal ideation. In this case, one might choose to regress the person-specific estimates of asymptote (a_i) on those manifest variables, such as: $a_i \sim N(\beta_a \mathbf{x}_i, \sigma_a^2)$, where β_a is a set of regression weights corresponding to person-level covariates $\mathbf{x}_i$, and σ_a^2 captures residual variation. This step is useful for establishing that individual differences in latent features are meaningfully related to person characteristics, such as asymptotic performance in our case would typically be related to age or mild cognitive impairment status as a characteristic and not as an outcome.

In applied healthcare settings, a clinician might want to use the results on the latent cognitive markers extracted from the repeated assessments differently: they want to predict the probability of a clinical outcome, such as MCI status. This could be useful, for example, if they have access to remotely collected response time data but lack the resources or access needed to bring an individual back to the clinic for comprehensive neuropsychological testing for establishing MCI status. Alternatively, such prediction could be part of some continuous monitoring for dementia risk that is based on remote ambulatory testing on people's smartphones. To be able to offer these inferences, we will establish predictive links directly between our latent cognitive markers and MCI status, while also appropriately accounting for demographic characteristics. That is, while in many empirical studies MCI is used on the right-hand side of predictive equations—predicting sources of individual differences in the latent cognitive features on the left-hand side—here we want to treat it as an outcome.

To establish the usefulness of the cognitive features for prediction, our analysis should involve a structural model making this prediction. Let us denote the manifest covariates of participant i with the vector $\mathbf{x}_i$, their latent features with the vector $\mathbf{\Lambda}_i$, and call their key, clinically meaningful outcome z_i (i.e., something interesting to detect, e.g., MCI status). We can then write the structural model for a binary outcome variable z_i as

$$\begin{cases} \pi_i = \text{logistic}\left(\beta_0 + \mathbf{B}_\lambda \mathbf{\Lambda}_i + \mathbf{B}_\mathbf{x} \mathbf{x}_i\right) \\ z_i \sim \text{Bernoulli}\left(\pi_i\right), \end{cases}$$

in which $\mathbf{B}_\lambda$ and $\mathbf{B}_\mathbf{x}$ are the vectors of regression weights that apply to the latent and manifest predictor vectors, respectively.

For our running example, we may try to predict MCI status from our latent cognitive markers while accounting for manifest demographic variables such as age, sex, education level and racial and ethnic differences. We would then choose

$$\begin{cases} \mathbf{B}_\lambda \mathbf{\Lambda}_i = \beta_a a_i + \beta_g g_i + \beta_r r_i + \beta_\varepsilon \varepsilon_i \\ \mathbf{B}_\mathbf{x} \mathbf{x}_i = \beta_{age} \text{Age}_i + \beta_{sex} \text{Sex}_i + \beta_{edu} \text{Edu}_i + \beta_{rac} \text{Rac}_i + \beta_{eth} \text{Eth}_i \\ \pi_i = \text{logistic}\left(\beta_0 + \mathbf{B}_\lambda \mathbf{\Lambda}_i + \mathbf{B}_\mathbf{x} \mathbf{x}_i\right) \\ \text{MCI}_i \sim \text{Bernoulli}\left(\pi_i\right), \end{cases} \quad [2]$$

where MCI_i is person i's MCI status, being predicted by a logistically transformed linear combination of predictors on the right hand side, namely β_0 denoting the intercept, coefficients β_a, β_g, β_r, and β_ε capturing the effect of our latent predictors a_i (asymptote), g_i (gain), r_i (learning rate) and ε_i (intra-individual variability), respectively, and coefficients β_{age} (Age), β_{sex} (Sex), β_{edu} (years of education, abbreviated to 'Edu'), β_{rac} (race, abbreviated to 'Rac'), and β_{eth} (ethnicity, abbreviated to 'Eth') quantifying the effect of their associated manifest predictors. The deterministic parameter π_i will be referred to as the "MCI risk," as it is the model-inferred probability that participant i has MCI.

Inference with the joint Bayesian multilevel model . The latent cognitive markers in the process model are estimated with uncertainty. We join the two component models in a multilevel Bayesian framework so that the uncertainty in prediction from different error sources is propagated in a statistically sound manner (Boehm, Marsman, Matzke, & Wagenmakers, 2018; Etz & Vandekerckhove, 2018; Wagenmakers et al., 2018). This approach allows for simultaneous estimation of cognitive markers, all regression coefficients, and variance components. The model is illustrated as a directed acyclic graph in Figure 2 (see, e.g., Lee & Wagenmakers, 2014, for more on the graphical model formalism). We can see there that the latent parameters simultaneously inform the cognitive performance data and the MCI status data. Multiple arrows emanating from each latent predictor is critical to the POP model, as the parameters must be constrained by the behavioral data in order to be identifiable and usable as predictors in the structural component model.

For our running example, this means specifying prior and hyperprior distributions for all parameters. For the process model parameters, we pool the person-level estimates via group-(population)level distributions:

$$\begin{cases} a_i \sim N(\mu_a, \sigma_a^2), & g_i \sim N(\mu_g, \sigma_g^2), \\ r_i \sim N(\mu_r, \sigma_r^2), & \varepsilon_i \sim N(\mu_\varepsilon, \sigma_\varepsilon^2). \end{cases} \quad [3]$$

For the hyperparameters, we typically choose non-informative priors that will not bias the estimation. On the means, we choose $\mu_\cdot \sim N(0, 1)$, and on the standard deviations, we set $\sigma_\cdot \sim N_+(0, 1)$ (the positive half-normal distribution). For the regression coefficients, we set similarly non-informative priors: $\beta_c \sim N(0, 1)$, where the c subscript stands in for 0 (intercept), a, g, r, ε (latent predictors), Age, Sex, Edu, Rac, and Eth (manifest predictors) in our current example. These prior settings assume that all covariates are standardized or dummy-coded.

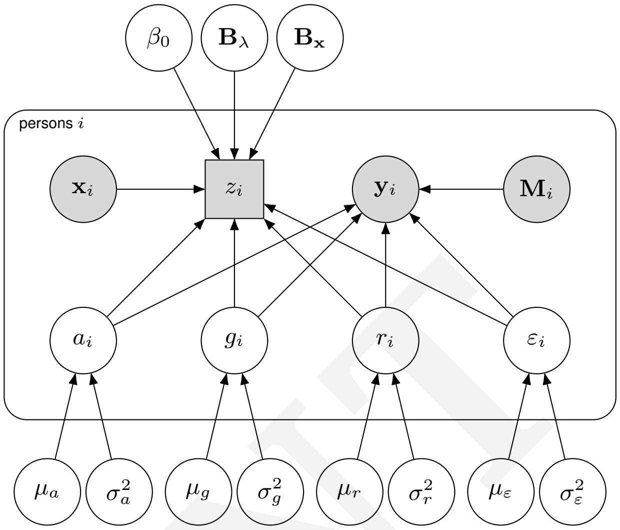

Fig. 2. A graphical model representation of our predictive model. In the graphical model formalism, unshaded nodes indicate parameters, shaded nodes indicate observed data, and square nodes indicate discrete values. Nodes that receive an arrow are partly determined by the node where the arrow originates. Plates indicate repetitions of their contents. The figure shows the four parameters of the model of practice: a_i, g_i, r_i, and ε_i. Each of these latent parameters is person-specific and is used to predict the behavioral data $\mathbf{y}_i$, which is the vector of response times provided by person i. $\mathbf{y}_i$ is additionally informed by the data node $\mathbf{M}_i$ (a vector of measurement times, or more generally information about the data collection design). The latent parameters are then used again to predict z_i, the mild cognitive impairment diagnosis. In that prediction, there is an intercept β_0, the latent variables have regression weights contained in $\mathbf{B}_\mathbf{\Lambda}$, and additional manifest variables $\mathbf{x}_i$ have regression weights contained in $\mathbf{B}_\mathbf{x}$.

The Bayesian implementation ensures that even when limited data is available (i.e., not much power to detect an effect), the resulting inferences are correct given that amount of data (but possibly with high posterior uncertainty; Wagenmakers et al., 2018). It also allows for non-binary statistical inference: Given the current data, we can easily express how much evidence we have for an effect. In our example, we might express the amount of evidence the data provide in favor of the predictive power of a particular predictor for MCI.

Application: Predicting mild cognitive impairment in the Einstein Aging Study

Study Design. The Einstein Aging Study (EAS) is an ongoing longitudinal research project examining risk factors for MCI and dementia. Participants of the EAS, all English-speaking, ambulatory residents of Bronx County, NY, aged 70 and above, were enlisted from local registered voting lists. The study was approved by the Albert Einstein College of Medicine Institutional Review Board and all participants gave written informed consent.

The latest analysis includes data from 316 participants. The average age of the sample at the outset of the study was 77.54 years, with a standard deviation of 4.98 years, and 67% were female ($n = 105$ male, and $n = 211$ female). The participant pool of the study reflected a diverse mix of racial and ethnic backgrounds, with 46.2% ($n = 146$) identifying as non-Hispanic Whites, 39.9% ($n = 126$) as non-Hispanic Blacks, 9.8% ($n = 31$) as Hispanic

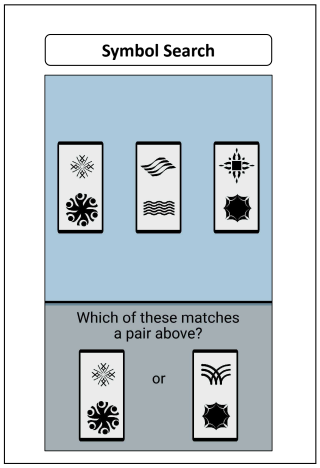

Fig. 3. Illustration of the Symbol search cognitive task from the Einstein Aging Study.

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 average educational level of the sample was 15.09 (SD = 3.55) years. Utilizing the Uniform Data System (UDS) neuropsychological assessment battery supplemented with the Free and Cued Selective Reminding Test (Katz et al., 2021) and Jak-Bondi criteria (Jak et al., 2009), 29.1% ($n = 92$) of participants were classified as having MCI at the study's baseline.

The EAS utilizes a measurement burst design, combining frequent ecological momentary assessments with in-clinic neuropsychological tests and demographic data collection. Over a 16-day period, participants engage in six short sessions daily—each lasting no more than five minutes—using smartphones provided by the study. These sessions, conducted during usual waking hours across various daily settings, consist of cognitive tasks ('brain games') and brief surveys on their immediate experiences, though the latter is not part of the current study's analysis. Four out of the six daily sessions are prompted at intervals of about 3.5 hours by random beeps throughout the day, while the first and last sessions are initiated by the participants themselves. In this study, numerous cognitive domains are assessed, but the current focus is on response time (RT) data from the Symbol Search task, which evaluates processing speed. The analysis focused on the daily mean response times.

Cognitive assessments with the Symbol Search task. In the current study, the Symbol Search task, depicted in Figure 3, is used to assess processing speed. During each trial, participants were presented with three pairs of symbols at the top of the screen and two pairs at the bottom. The task required participants to quickly and accurately identify which one of the bottom pairs matched one of the top three pairs. Each session consisted of 11 trials. For analysis, we compiled daily aggregates of the reaction times (RTs) for correctly matched trials and examined them using the Bayesian Exponential Model.

Demographic variables. Participant demographic data were obtained through questionnaires. For the purposes of the present analysis, we used the following demographic variables: age (expressed in years and standardized), sex (categorized as male or female based on participants' self-declared sex, with male serving as the reference category), educational attainment (measured in total years of schooling and standardized), and ethnicity (classified as Caucasian, African American, Hispanic White, Hispanic Black, Asian, or Other). Wang et al. (2021) provide detailed discussion of these demographic variables and their relevance for predicting MCI status.

Mild cognitive impairment status. Each participant was subjected to a comprehensive neuropsychological evaluation to determine their cognitive status. This was an in-clinic assessment and encompassed tests for memory, executive function, attention, language, and visuospatial skills (Katz et al., 2021). The criteria for MCI classification adhered to the Jak-Bondi criteria (Jak et al., 2009).

Results.

Implementation. We implemented the model (using Eqs. 1–3) in Stan (B. Carpenter et al., 2017), which can be accessed from R (R Core Team, 2022) through the `rstan` package (Stan Development Team, 2023). Code is available at the OSF page of the study: `https://osf.io/4qpxs/`. We analyzed the EAS data by running 4 parallel chains, 2500 warm-up plus 2500 posterior samples per chain, for a final posterior sample size of 10,000. We did not find any problems with convergence based on the diagnostic criterion $\hat{R}$ (Gelman et al., 2013; all $\hat{R} < 1.03$) and visual check of the sample chains. We checked the quality of the sampling by calculating effective sample size (the proportion of samples that can be considered as non-correlated draws in the posterior), which showed good sampling quality (above 300 for 99.8% parameters, and above 1000 for 98% of the parameters, including all the key hierarchical parameters). Analysis took less than 20 minutes on a MacPro laptop.

Model fit. We checked model fit first with visual inspection of the correspondence between each individual's observed data and their model-predicted learning curve (all generated plots are available on the OSF page of the project). Then we used those data and curves to compute the proportion of variance in the data that is explained by the process model (akin to an R^2 statistic), which was .83. Both methods showed acceptable fit. However, both of these methods focused only on the fit of the process model to the observed data, and not on the prediction of MCI with the logistic regression component from Eqs. 2 (for which, see the Predictive Accuracy section).

Parameter estimates. The top part of Table 1 shows the results from the logistic regression coefficients linking the latent features and manifest covariates to MCI status. The first column shows the name of the predictors/covariates, the second the posterior mean as a point estimate for their corresponding regression coefficient, and the last two columns display the directional probability. Instead of using a binary rule, such as excluding zero from a 95% interval, this allows for a graded approach to quantify support for effects. In this application, we will consider coefficients with directional probabilities between 91% and 95% as possibly credible effects, between 95% and 99% as likely credible effects, and over 99% strongly credible effects. The intercept is the first one of these, with its entire posterior mass in the negative range, suggesting an overall larger probability of not having MCI than having MCI in this sample, which makes sense given that only about one-third of the whole sample had MCI. From our latent cognitive markers, two showed credible links to MCI status: higher asymptotic performance (i.e., slower peak performance response time), and slower learning rate both predicted MCI. Regarding our manifest predictors, being Black tended to correspond to higher probability for MCI.

Table 1. Regression coefficient estimates and their corresponding directional probabilities and estimates of the hierarchical parameters of the latent predictors with 95% credible intervals.

Predictors	mean	P(negative)	P(positive)
Intercept	-2.8047	1.0000	0.0000
Asymptote	0.6362	0.0027	0.9973
Learning Rate	-1.4253	0.9923	0.0077
Intra-individual variability	0.8513	0.0671	0.9329
Gain	-0.0339	0.5790	0.4210
Age (standardized)	0.1929	0.0794	0.9206
Sex (1: male)	-0.1975	0.7606	0.2394
Years of education (standardized)	0.1907	0.0985	0.9015
Race (1: Black)	0.4855	0.0469	0.9531
Ethnicity (1: Hispanic)	0.3031	0.2391	0.7609

Group-level summaries	mean	PCI	PCI
		2.50%	97.50%
Asymptote Mean	2.9533	2.8468	3.0606
Asymptote SD	0.9053	0.8170	1.0056
Learning Rate Mean	0.5367	0.4686	0.6106
Learning Rate SD	0.2989	0.2462	0.3604
Intra-individual variability Mean	0.7658	0.7145	0.8178
Intra-individual variability SD	0.4436	0.4011	0.4903
Gain Mean	1.8067	1.6276	1.9913
Gain SD	1.1512	1.0073	1.3037

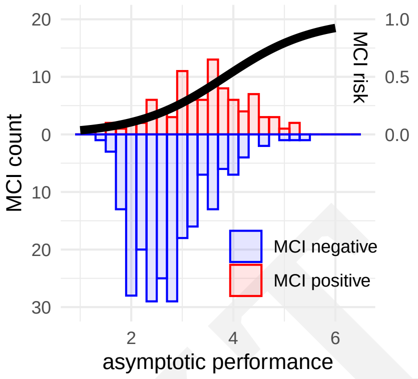

Fig. 4. An illustration of the nonlinear predictions made by the partially observable predictor model. On the horizontal axis is the person-specific estimate of the asymptote parameter a_i. The downward-pointing (blue) histogram shows the number of participants in the MCI-negative group as a function of a_i, and the upward-pointing (red) histogram shows the corresponding number of participants in the MCI-positive group. The red histogram is somewhat to the right of the blue histogram, indicating that participants in the MCI-positive group tend to have higher asymptotes a_i. The thick black curve shows the MCI risk for participants with average (or baseline) values on all predictors except a_i. As expected, the curve is near 0 in the range of a_i where most participants have negative MCI status, rises to above 0.5 at $a_i \approx 4$, but does not approach 1 anywhere in the realized range of a_i.

There was weak evidence that more intra-individual variability (i.e., inconsistency in performance) and older age were also predictive for MCI status. The bottom part of Table 1 additionally shows estimates of the hierarchical parameters of the latent predictors with corresponding 95% credible intervals.

Interpretation of parameters. Using the posterior distributions of these parameter estimates in combination with Equation 2, we can calculate how MCI risk changes as a function of our cognitive markers. For example, if we consider a participant who has average values on everything (including latent and manifest predictors, and is male, White, and non-Hispanic for the dummies), their probability of MCI is .25. However, if this individual is 1 SD higher than the average regarding their asymptotic performance (e.g., slower reaction time), this probability goes up to .37, and if they are Black, it increases further to .49. Additionally, if this individual is also 1 SD lower in learning, then the probability increases to .59. Figure 4 further illustrates the effect of the latent variable a_i, which captures the asymptotic performance of a participant i. The two participant populations (histograms of MCI positive and MCI negative participant numbers) visibly separate along the horizontal axis, where the latent variable is plotted. An S-shaped curve connects a_i to participant i's MCI risk (denoted π_i in Eq. 2) and illustrates the effect of a_i on our MCI prediction for a participant whose other predictors are otherwise at baseline (for Sex, Race, and Ethnicity) or at the population average (for others).

Predictive accuracy. The goal of our POP model approach is to decompose MCI risk into constituent components with their own psychological interpretability – a mix of manifest variables like age and sex with latent variables like asymptotic ("peak") performance. However, the interpretability of these components comes at a cost: the added complexity of the model may harm its predictive performance ("overfitting;" Hastie, Tibshirani, & Friedman, 2009), and we may not be willing to trade much predictive accuracy for the benefit of interpretability. For this reason, we evaluate the predictive performance of our proposed model and compare it against two more conventional models: one that uses only manifest predictors, and one simplified POP model that uses behavioral data but no process model to aid in interpretability.

Evaluation metrics. To evaluate predictive performance, we calculated the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, which summarizes the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Swets, 1988). The AUC is a preferred statistic especially when dealing with imbalanced datasets in which accuracy alone may be misleading.

Cross-validation. We calculated the AUC via a ten-fold cross-validation procedure (Hastie et al., 2009). From the original data set ($N = 316$) we first created ten subsets. We stratified the subsets such that, like the full data set, each subset contained about 30% MCI positive participants. We then fit the model ten times, each time holding out the MCI status of a different subset of participants. We estimated their latent process parameters from their cognitive task data and combined those with their manifest predictors to obtain a person-specific predicted MCI risk π_i^{pred} for each holdout participant. We then created a ROC curve by varying the critical value of π_i^{pred} by which we categorized participants as MCI negative or positive, and we computed a confidence interval around the ROC curve using a bootstrap procedure provided by R's pROC package (J. Carpenter & Bithell, 2000; Robin et al., 2011).

Comparator models. For comparison, we then similarly calculated ROC curves for a model using only the manifest predictors (the "Manifest only" model) and a POP model that combines manifest predictors with a simplified generative model of RT (describing RT only in terms of mean and standard deviation; the "Manifest + latent descriptors" model).

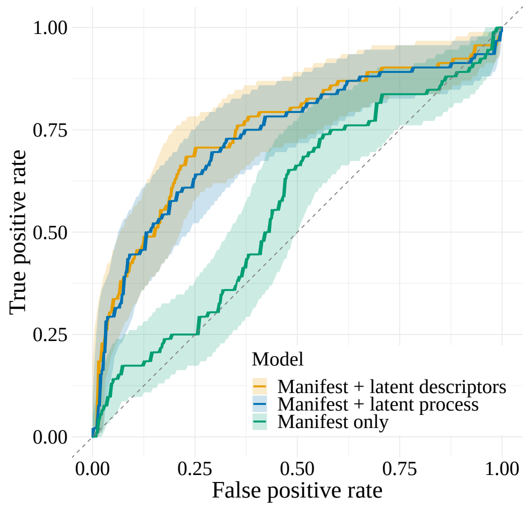

Fig. 5. Receiver operating characteristic (ROC) curve of two partially observable predictor models, compared to a model that only uses manifest predictors. On the vertical axis is the true positive rate (correct predictions of positive MCI status) and on the horizontal axis is the false positive rate (incorrect predictions of positive MCI status). The area under the curve (AUC) is a concise summary of the predictive ability of the model. The AUC is lower for the "Manifest only" model but indistinguishable for the "Manifest + latent descriptors" and "Manifest + latent process" models. Shaded bands around the ROC curve indicate a 95% confidence interval obtained through a bootstrap procedure (J. Carpenter & Bithell, 2000; Robin et al., 2011).

Results. ROC curves for the three models are shown in Figure 5. The AUC of the "Manifest + latent process parameters" model was .7346 (95% CI: [0.6680, 0.8013]). The AUC of the "Manifest only" model was .5640 (95% CI: [0.4938, 0.6342]), with its confidence interval including .5, and the AUC of the "Manifest + latent descriptors" model was .7371 (95% CI: [0.6717, 0.8026]).

Discussion. The results of this cross-validation exercise led to three conclusions. First, the lower AUC for the "Manifest only" model speaks to the predictive utility of behavioral data above and beyond demographics. The AUC for this model was close to .5, suggesting that predicting MCI status based solely on demographic information yields accuracy nearly equivalent to random guessing. Second, the observation that the AUCs and ROCs for the "Manifest + latent process parameters" and "Manifest + latent descriptors" were indistinguishable assuages any concerns of overfitting by the more complex process model. Third, the AUCs of the POP models are fair but also leave room for improvement (see, e.g., the examples in Swets, 1988). Future work focusing on predictive accuracy could use POP models to factor in additional sources of evidence, possibly from a variety of tasks in a battery. However, for the current data we conclude from this analysis that the "Manifest + latent process parameters" model improves interpretability with no concomitant reduction in predictive accuracy.

Conclusions and discussion

There is a growing consensus that 40% of dementia cases are due to modifiable risk factors (Lee et al., 2022). Understanding the early signs of subtle cognitive decline, for example in preclinical ADRD, can open up new possibilities for secondary prevention and monitoring. These subtle early cognitive changes are most likely related to specific subprocesses that we need to identify. We have introduced partially observable predictor (POP) models, an approach that combines observable demographic variables with unobservable predictors derived from behavioral data. We demonstrate how this approach may be applied to test the role of cognitive process parameters for early detection. POP models let us discern which underlying psychological components are relevant for predicting clinically meaningful outcomes, offering a clearer picture on early-stage neuropsychological impairments.

Conclusions from the Einstein Aging Study. Using the Einstein Aging Study (EAS) data set, we have demonstrated that data from repeated cognitive measures improved prediction accuracy when compared to a model with only manifest variables. Furthermore, we were able to decompose the contribution of these cognitive measures using a cognitive process model, which allowed us to compare the individual contribution of interpretable cognitive subprocesses. For the processing speed task in the EAS, the learning rate and peak performance turned out to be important for prediction of mild cognitive impairment.

Limitations of the example data set. Even though our predictive model pooled information from multiple sources, total prediction accuracy was lower than in some previously published studies (e.g., Oh, Kim, & Lee, 2024; Yan, Zhang, & Chen, 2023). This was true independently of the application of the POP model (i.e., it was also true when using only the manifest predictors), leading us to believe that the MCI categorization used in the EAS data (Chang et al., 2024) was noisier than other methods (Devlin et al., 2022). Indeed, in longitudinal data with this categorization method, participants even occasionally changed status from positive back to negative over time.

Potential for future uptake. Since the response time data that we used to derive our cognitive markers was collected in ambulatory settings—specifically, via smartphones—the approach facilitates easy screening and monitoring of mild cognitive impairment risks. These measures represent naturalistic, real-time functioning. While neuropsychological evaluations are widely regarded as the definitive method for identifying cognitive deficits, they are extensive and generally need to be conducted in person, restricting their broad applicability in large-scale research and clinical studies. The EAS data were collected with a cognitive task that could easily be collected remotely, multiple times per year, representing a low barrier for entry for underrepresented populations who might not have easy access to clinicians.

We are cautiously optimistic about the broader appeal and potential for uptake of our proposed approach. Given that cognitive researchers could use POP models to incorporate many different process models—whichever is more appropriate for the behavioral data at hand—into a one-step predictive model, we anticipate interest. The key barrier to uptake is the technical challenge involved in the implementation of a Bayesian multilevel model in Stan. However, Bayesian methods and models are no longer the obscure niche skill they once were – the increase in relevant course books and tutorials (Lee & Wagenmakers, 2014; McElreath, 2020; Vandekerckhove, Rouder, & Kruschke, 2018; Wagenmakers et al., 2018) strongly suggest that Bayesian cognitive modeling is becoming a leading paradigm in computational cognitive science that is suitable for translation to clinical science (Huys, Maia, & Frank, 2016; Maia & Frank, 2011). Additionally, we have made all of our data analysis code (Stan/R scripts and an associated Dockerfile) freely available via OSF. To facilitate adoption in this field even more, it would be optimal to have a user-friendly online platform that integrates data collection with the delivery of analytical results to clinicians. Although we do not currently have such a platform, building such tools has been an NIH funding priority (e.g., National Institute on Aging, 2022), and recent progress (e.g., Hakun, Elbich, Roque, Yabiku, & Sliwinski, 2024) gives us confidence in these becoming a reality.

Future work. Further studies that use POP models in different data sets might identify other subprocesses extracted from cognitive tasks capturing performance on a different cognitive domain, and a combination of information from multiple multi-domain tasks could lead to increased predictive accuracy, but more importantly to greater understanding that could inform targeted intervention.

Secondly, while we used concurrent MCI status in the current application, it is possible to gather data on which non-MCI participants at baseline develop MCI in the future ('incident' MCI). In future work, we will use this approach for prediction in the epidemiological sense, with a focus on predicting an outcome in the future based on information in the present.

Finally, we note that while we focused on cognitive models, the presented approach can be applied across various fields. In computational psychiatry, for instance, a comparable strategy is used to distill latent emotional dynamics, which can then inform predictive models for critical clinical outcomes like suicide risk.

References

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model–based approaches* (pp. 71–93). Washington, DC, US: American Psychological Association. doi: 10.1037/12074–004

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018, August). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods*, *50*(4), 1614–1631. doi: 10.3758/s13428-018-1054-3

Browning, M., Behrens, T. E. J., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature neuroscience*, *18*, 590–596.

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).

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, *19*, 1141–1164. doi: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Chang, K. H., Katz, M. J., Qin, J., Wang, C., Lipton, R. L., Byrd, D. A., & Rabin, L. A. (2024). *Comparing predictive validity for four MCI classifications in demographically-diverse community-dwelling individuals: Results from the Einstein Aging Study (EAS).*

Charpentier, C., Aylward, J., Roiser, J., & Robinson, O. (2016, 12). Enhanced risk aversion, but not loss aversion, in unmedicated pathological anxiety. *Biological Psychiatry*, *81*. doi: 10.1016/j.biopsych.2016.12.010

Devlin, K. N., Brennan, L., Saad, L., Giovannetti, T., Hamilton, R. H., Wolk, D. A., ... Mechanic-Hamilton, D. (2022, January). Diagnosing mild cognitive impairment among racially diverse older adults: Comparison of consensus, actuarial, and statistical methods. *Journal of Alzheimer's Disease*, *85*(2), 627–644. doi: 10.3233/jad-210455

Dzierzewski, J., Marsiske, M., Aiken Morgan, A., Buman, M., Giacobbi, P., Roberts, B., & McCrae, C. (2013, September 1). Cognitive inconsistency and practice-related learning in older adults. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *26*(3), 173–184. doi: 10.1024/1662-9647/a000096

Erdfelder, E., Hilbig, B., Auer, T.-S., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009, 01). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 108-124.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. doi: 10.3758/s13423-017-1262-3

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, Third edition*. Boca Raton (FL): Chapman & Hall/CRC.

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. doi: 10.31234/osf.io/34ux5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY, USA: Springer. doi: 10.1007/978-0-387-84858-7

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207. doi: 10.3758/bf03212979

Hernaus, D., Gold, J. M., Waltz, J. A., & Frank, M. J. (2018). Impaired expected value computations coupled with overreliance on stimulus-response learning in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(11), 916–926. doi: 10.1016/j.bpsc.2018.03.014

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. doi: 10.1038/nn.4238

Jak, A., Bondi, M., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D., & Delis, D. (2009, 06). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry*, *17*, 368-75. doi: 10.1097/JGP.0b013e31819431d5

Katz, M. J., Wang, C., Nester, C. O., Derby, C. A., Zimmerman, M. E., Lipton, R. B., ... Rabin, L. A. (2021, January). T-moca: A valid phone screen for cognitive impairment in diverse community samples. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *13*(1). doi: 10.1002/dad2.12144

Lee, M., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lee, M., Whitsel, E., Avery, C., Hughes, T. M., Griswold, M. E., Sedaghat, S., ... Lutsey, P. L. (2022, July). Variation in population attributable fraction of dementia associated with potentially modifiable risk factors by race and ethnicity in the us. *JAMA Network Open*, *5*(7), e2219672. doi: 10.1001/jamanetworkopen.2022.19672

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. doi: 10.1038/nn.2723

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in r and stan* (2nd ed.). New York, NY, USA: Chapman and Hall/CRC. doi: 10.1201/9780429029608

National Institute on Aging. (2022, July 7). *Par-22-213l Complex Integrated Multi-Component Projects in Aging Research (U19 Clinical Trial Optional).* `https://grants.nih.gov/grants/guide/pa-files/PAR-22-213.html`.

Oh, J., Kim, S., & Lee, H. (2024). Using sociodemographic variables for the prediction of mild cognitive impairment in an aging population: A machine learning approach. *JMIR Medical Informatics*, *12*(1), e59396.

Oravecz, Z., Faust, K., & Batchelder, W. (2014). An extended cultural consensus theory model to account for cognitive processes in decision making in social surveys. *Sociological Methodology*, *44*, 185-228. doi: 10.1177/0081175014529767

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. *The importance of cognitive practice effects in aging neuroscience*, *16648714*, 128.

Oravecz, Z., Vandekerckhove, J., Hakun, J., Kim, S., Katz, M., Wang, C., . . . Sliwinski, M. (2025). Computational phenotyping of cognitive decline with retest learning. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*. doi: 10.1093/geronb/gbaf030

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Ratcliff, R., Scharre, D. W., & McKoon, G. (2022). Discriminating memory disordered patients from controls using diffusion model parameters from recognition memory. *Journal of Experimental Psychology: General*, *151*(6), 1377–1393. doi: 10.1037/xge0001133

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, *12*, 77.

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

Stier, A. J., Sajjadi, S., Karimi, F., Bettencourt, L. M. A., & Berman, M. G. (2024). Implicit racial biases are lower in more populous more diverse and less segregated us cities. *Nature Communications*, *15*(1). doi: 10.1038/s41467-024-45013-8

Swets, J. A. (1988, June). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. doi: 10.1126/science.3287615

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*(3), 415.

Thompson, L., Harrington, K., Roque, N., Strenger, J., Correia, S., Jones, R., . . . Sliwinski, M. (2022, 04). 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*. doi: 10.1002/dad2.12283

Vandekerckhove, J., Rouder, J., & Kruschke, J. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4. doi: 10.3758/s13423-018-1443-8

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. (2011). Hierarchical diffusion models for two-choice response times. *Psychological methods*, *16*(1), 44.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: 10.3758/s13423-017-1343-3

Wang, C., Katz, M., Chang, K., Qin, J., Lipton, R., Zwerling, J., . . . Abner, E. (2021). Udsnb 3.0 neuropsychological test norms in older adults from a diverse community: Results from the einstein aging study (eas). *Journal of Alzheimer's Disease*, *83*(4), 1665–1678. doi: 10.3233/JAD-210538

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the expectancy valence model of the iowa gambling task. *Journal of Mathematical Psychology*, *54*, 14-27. doi: 10.1016/j.jmp.2008.12.001

Yan, L., Zhang, Y., & Chen, W. (2023). Predicting current mild cognitive impairment in a high stroke-risk population using demographic and clinical factors. *Frontiers in Aging Neuroscience*, *15*, 1180351.

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science*, *16*(12), 973-978. (PMID: 16313662) doi: 10.1111/j.1467-9280.2005.01646.x

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Interested collaborators are asked to complete a concept proposal form (details for potential project, paper, or abstract) to be reviewed and forwarded to the Einstein Aging Study Steering Committee for consideration. Requests to access these datasets should be directed to MK, MPH at mindy.katz@einsteinmed.edu. For additional information on data sharing requests for the Einstein Aging Study, see `https://einsteinmed.edu/departments/neurology/clinical-research-program/eas/data-sharing.aspx`.

Acknowledgements

ZO and MS were supported by the NIH (R01AG074208, R56AG074208 and U2CAG060408). MK received funding from the NIH (NIA 2 P01 AG03949), the Leonard and Sylvia Marx Foundation, and the Czap Foundation. JV was supported by the National Science Foundation (#1658303, #1850849, and #2051186).

Author contributions statement

All authors contributed to the final draft. Correspondence concerning this article should be addressed to Zita Oravecz (zita@psu.edu).

Competing interests

The authors declare no competing interests.