Journal of Mathematical Psychology 55 (2011) 47–56

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Understanding memory impairment with memory models and hierarchical Bayesian analysis

James P. Pooley [a,*], Michael D. Lee [a], William R. Shankle [a,b]

[a] Department of Cognitive Sciences, University of California, Irvine, United States
[b] Medical Care Corporation, University of California, Irvine, United States

ARTICLE INFO

Article history:
Received 18 February 2010
Received in revised form
26 July 2010
Available online 23 October 2010

Keywords:
Alzheimer's disease
Cognitive psychometrics
Dementia
Human memory
Measurement models

ABSTRACT

The study of human episodic memory is a topic that interests cognitive and mathematical psychologists as well as clinicians interested in the diagnosis and assessment of Alzheimer's disease and related disorders (ADRD). In this paper, we use simple cognitive models for the recognition and recall tasks typically applied in clinical assessments of ADRD to study memory performance in ADRD patients. Our models make use of hierarchical Bayesian methods as a way to model individual differences in patient performance and to facilitate the modeling of performance changes that occur during multiple recall tasks. We show how the models are able to account for different aspects of patient performance, and also discuss some of the predictive capabilities of the model. We conclude with a discussion on the scope to improve on our results by discussing the link between memory theory in psychology and clinical practice.

1. Introduction

Much recent work in cognitive and mathematical psychology has focused on the application of psychological models to clinical data (e.g., Neufeld, 2007). In areas ranging from reinforcement learning models of decision making in substance abuse populations (e.g., Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010) to response time modeling in elderly populations (e.g., Ratcliff, Thapar, & McKoon, 2001), useful links are being made between the more theoretical and the applied sides of psychology.

One cognitive domain in particular that has long attracted the interests of both psychological and clinical researchers is episodic memory, a type of memory (or memory system) for personally experienced events (e.g., Tulving, 2002). Episodic memory affects many aspects of an individual's daily life and is particularly important for understanding a variety of neurological disorders that are associated with dementia. Collectively referred to as Alzheimer's disease and related disorders (ADRD), these disorders are characterized by a variety of cognitive deficits. However, the effects of ADRD on episodic memory are particularly severe (e.g., Hodges, 2000). In fact, for Alzheimer's disease, as well as for many of the other disorders that comprise ADRD, differences in the severity of episodic memory degradation best distinguish adults who are aging normally from those affected with dementia on the basis of behavior alone (e.g., Locascio, Growdon, & Corkin, 1995).

Although great advances in our understanding of the neuropathological basis of ADRD of have been made since the initial description of Alzheimer's disease (Hodges, 2006), comparatively little progress has been made toward a rigorous description of the actual memory deficits these disorders cause. Behavioral memory data from ADRD patients possess a rich structure in the sense that the data are obtained from patients who can be clustered into distinct groups based on, for example, the severity of their ADRD impairment. In practice, this structure is often ignored for simplicity. Related to this methodological issue is a more philosophical one. Specifically, it seems desirable to use methods and models that deal directly with the memory concepts that one is interested in, and it is unclear how well off-the-shelf statistical models can accomplish this. Fortunately, there is a clear path to applying what has been learned from psychological research to the problem of memory assessment in ADRD patients.

The key link is that some of the most important ADRD assessment tools standardly used in clinical practice are exactly those memory tasks that are studied by memory researchers in cognitive and mathematical psychology for years. These fields, in turn, have developed numerous mathematical and computational models of memory based on behavioral data from these tasks. The utility of these cognitive models is that they allow researchers to characterize memory performance in terms of psychologically meaningful variables (e.g., recognition bias) rather than standard statistical variables (e.g., regression coefficients). Thus, we feel that the use of memory models, together with the use of hierarchical Bayesian methods to connect these models to the memory data of ADRD patients, can potentially be a useful complement to research on the neuropathology of ADRD.

* Corresponding address: Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, 92697-5100, United States.
E-mail address: jpooley@uci.edu (J.P. Pooley).

0022-2496/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2010.08.003

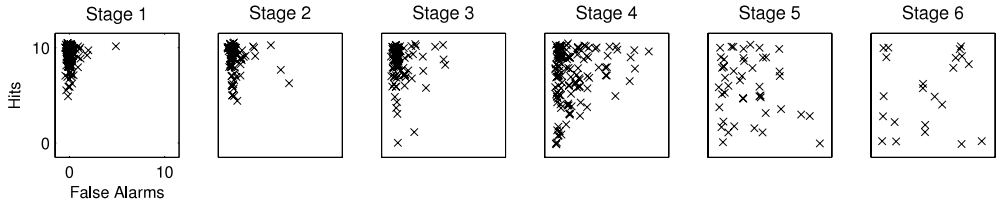

Fig. 1. Recognition data from the clinical ADRD assessments. Each panel corresponds to a FAST stage. The crosses in each panel correspond to a patient and show the hit and false alarm counts produced by this patient.

In this paper, we present such an approach to understanding the episodic memory deficits associated with ADRD, all within a hierarchical Bayesian framework using cognitive models. This paper is organized as follows: first, we give an overview of a relatively large clinical data set and the assessment protocols that underly the memory data it contains. Next, we outline the two memory models that we will use to account for the data, as well as the hierarchical Bayesian methods and related mathematical tools we use to connect these models to the data. Following this, we present the results of our analysis and discuss what they have to say about the episodic memory performance in ADRD at both a group and individual patient level. We conclude with a discussion of the implications of these results for both clinical practice and basic psychological research.

2. Clinical data

Our memory data comprise a subset of a large clinical ADRD database. This database contains a wealth of information on ADRD patients (and often on their caregivers as well) who have visited neurology clinics located in the United States for routine dementia screening and assessment, including potentially relevant demographic information and information concerning personal medical history. In addition to this medical information (and more likely to be of interest to psychologists), this database also contains the results of various psychological tasks that are administered as part of the cognitive portion of these dementia assessments. Of these psychological tasks, however, we focus exclusively on a recognition memory task and sequence of four free recall memory tasks.

Stimuli for each memory task consisted of words selected from the CERAD (Consortium to Establish a Registry for Alzheimer's Disease) word list (Morris, Mohs, & Rogers, 1988), which serves as the basis for the neuropsychological portion of many ADRD assessments. These words, which included a mixture of common nouns, were chosen with the goal of minimizing the degree to which they influence patient performance on the memory tasks. Based on these stimuli, the following assessment protocols were used to obtain the memory data.

2.1. Recognition task and data

Each patient first completed a standard old/new recognition task. In this task, patients were presented with a list of 10 words to study. Following the presentation of this study list, the patients were presented with a test list consisting of these 10 studied words as well as 10 non-studied words, and the patients were instructed to indicate whether or not a given word on the test list was on the study list by responding OLD or NEW accordingly. Based on these dichotomous responses, the recognition data for each patient take the form of four counts: (1) *hits* (i.e., correct OLD responses to studied words), (2) *misses* (i.e., incorrect NEW responses to studied words), (3) *false alarms* (i.e., incorrect OLD responses to non-studied words), and (4) *correct rejections* (i.e., correct NEW responses to non-studied words). Since the numbers of studied and non-studied words are typically treated as known quantities that are part of the experimental design, there are redundancies in the above counts, and it is typical to consider only the hit and false alarm counts. These hit and false alarm counts for each of the 525 patients are displayed in Fig. 1.

2.2. Recall tasks and data

Following this recognition task, each patient completed a sequence of four free recall tasks: three immediate free recall tasks followed by one delayed free recall task. On the first three of these tasks, each patient was shown a list of 10 words to study. Following this study period, each patient was instructed to recall as many of the words as possible. Each patient saw the same 10 words in the same order on each task. Following these three immediate free recall tasks, and following the administration of a distractor task (an unrelated cognitive task administered as part of the dementia assessment), each patient was given a surprise delayed free recall task in which they were required to recall as many of the 10 words that comprised the previous study list as possible. The data for each patient on each task consist of a binary sequence that indicates whether or not the patient recalled the word at each of the serial positions on the study list. It is much more common, however, to work with data that have been averaged over patients, and show, at each serial position, the proportion of patients that recalled a given word. These averaged data are shown in Fig. 2.

2.3. FAST stage classifications

Independent of the performance of each patient on these memory tasks, a trained neurologist used the functional assessment staging test (FAST) to classify the severity of each patient's dementia. The FAST (Reisberg, 1988) is a diagnostic tool used by clinicians to track the progression of ADRD by classifying patients into one of the seven *stages* in terms of the severity of dementia. Our data set contains individuals classified in each stage except for FAST stage 7. At best, patients with a classification of FAST stage 7 can speak approximately six or seven words per day; at worst, these patients are unable to lift their heads. Consequently, no data from patients with this classification were included in our data set. Important characteristics of the FAST are summarized in Table 1, including the number of patients from each stage whose data we are modeling. In total, our data set contains the memory data and FAST stages for 525 patients.

3. Measurement models for recognition and recall

Cognitive and mathematical psychologists have developed numerous mathematical and computational models in order to account for performance on episodic memory tasks, ranging from abstract process models to neural network models that are intended to have some degree of biological plausibility (e.g., Norman, Detre, & Polyn, 2008). Each type of memory model has its merits, some of which make many of them unnecessary to use in our current application, where all we require are models whose parameters have psychological interpretations. For this reason, we adopt simple psychological measurement (cf. process) models (Batchelder, 1998). A key feature of these models is that they "…capture some of the psychologically important variables in a paradigm, but they are necessarily approximate and incomplete and are usually confined to particular paradigms" (Batchelder & Riefer, 1999, p. 58).

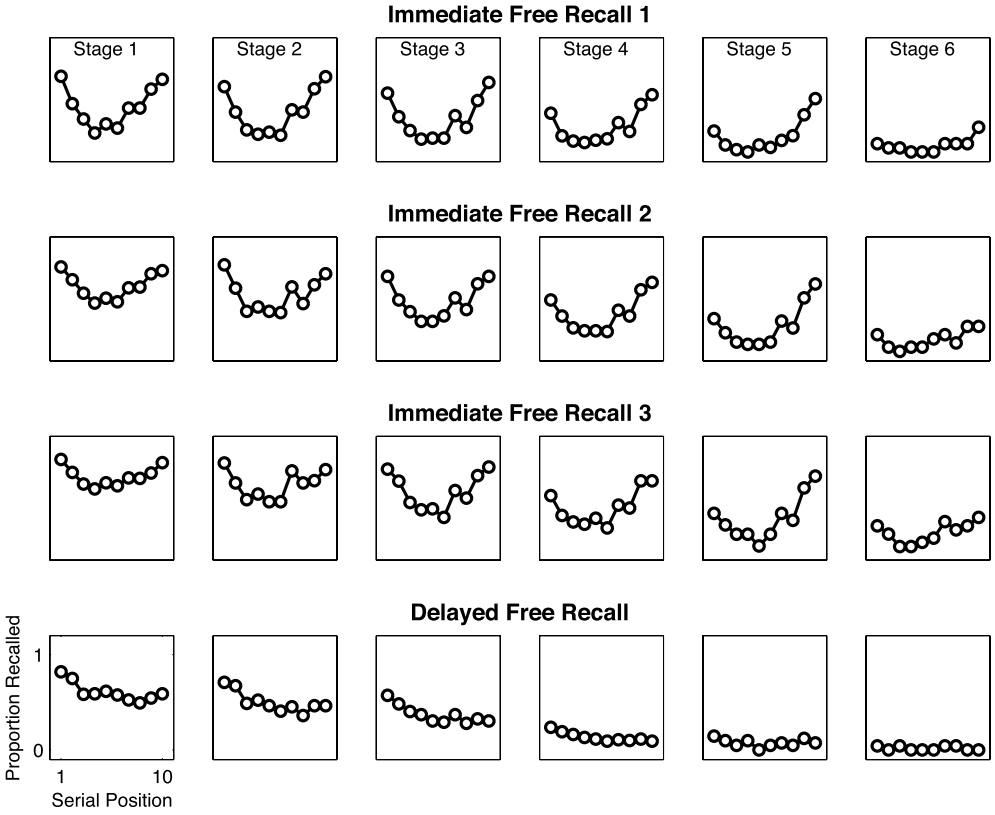

Fig. 2. Recall data from the clinical ADRD assessments. Each column corresponds to a FAST stage, and each row corresponds to one of the recall tasks. The data are averaged over patients and show, for each task, the proportion of words recalled at each of the serial positions on the study list.

Table 1
FAST stages.

Stage	Name	Characteristics	Number
1	Normal aging	No deficits whatsoever	156
2	Possible mild cognitive impairment	Subjective functional deficit	86
3	Mild cognitive impairment	Objective functional deficit interferes with a person's most complex tasks	89
4	Mild dementia	Troubles with bill paying, cooking, cleaning, traveling	130
5	Moderate dementia	Needs help selecting proper attire	41
6	Moderately severe dementia	Needs help bathing	23
7	Severe dementia	Can no longer hold up head	0

3.1. A SDT model for recognition

To account for the recognition data, we adopt the standard equal-variance Gaussian signal detection theory (SDT) model (e.g., Wickens, 2002, Chapter 2). In this model, an individual's memory representations for the studied and non-studied words are modeled as unit variance[1] Gaussian distributions over a unitary dimension of memory strength. The mean of the distractor distribution is arbitrarily set to 0, so the mean of the target distribution d' is the separation of the target and distractor distributions and so referred to as the *discriminability* of the items.

Due to the overlap of the target and distractor distributions at the point $d'/2$ on the memory strength continuum, an individual needs a decision strategy for relating subjective memory strength to decisions on the old/new recognition task. According to SDT, an individual accomplishes this by choosing a criterion level of memory strength λ above which the individual always responds OLD (i.e., "this word was on the study list") and below which the individual always responds NEW (i.e., "this word was not on the study list").

The point of overlap $d'/2$ of the target and distractor distributions is the optimal placement of the criterion in the sense that individuals with this criterion show no preference for one particular response, so responses made using this criterion are *unbiased*. However, it is often the case that an individual will be more or less concerned with one aspect of the old/new recognition task. In the parlance of SDT, the individual may choose to be *conservative* or *liberal* in deciding the placement of the criterion. As such, the distance $c = \lambda - d'/2$ between an individual's actual criterion and the unbiased criterion measures an the *response bias* of the individual, with $c > 0$ indicating that the individual has a conservative response bias, $c < 0$ indicating that the individual has a liberal response bias, and $c = 0$ indicating that the individual is unbiased in the placement of their criterion.

3.2. A two-factor model for recall

To account for the recall data, we developed a simple two-factor model with the goal of accounting for the serial position curve. Two-factor models have a history in the memory literature

[1] In current memory theory, it is more common to adopt an unequal-variance assumption in which the distribution for the studied words is 25% more variable than the distribution for the non-studied words (e.g., Ratcliff, Sheu, & Gronlund, 1992). Preliminary simulations using this unequal-variance assumption produced identical results to those presented in this paper, so we decided to to use the simpler, equal-variance SDT model in the current application.

(e.g., Henson, 1998), and provide a simple (if incomplete) method of accounting for salient properties of free recall data. The current model contains two parameters: a primacy parameter α that controls the probability of recalling a word presented near the beginning of the study list, and a recency parameter β that controls the probability of recalling a word presented near the end of the study list. These parameters can roughly be thought of as a form of memory strength, in the same way that d' in SDT is a measure of memory strength. However, we are *not* asserting that a two-store model (e.g., Raaijmakers & Shiffrin, 1981) is the mechanism responsible for memory performance.

In particular, the kth presented word in the list with N words has primacy recall probability α^k and recency recall probability β^{N-k+1} according to the model. These terms then combine multiplicatively to give the probability of recalling word k as $\theta_k = 1-(1-\alpha^k)(1-\beta^{N-k+1})$. We note that this model makes a number of unrealistic simplifying assumptions (e.g., there is no internal re-ordering of the study list by the patients and no mechanism for dealing with false recalls is provided), but is sufficient for our exploratory goals in the current application.

4. Hierarchical Bayesian analysis

The Bayesian approach to statistical inference and data analysis has a number of conceptual and practical advantages when compared to classical statistics (e.g., Gelman, Carlin, Stern, & Rubin, 2004). One such advantage is the relative ease with which hierarchical (or multilevel) modeling is accommodated, making it relatively easy to work with multiparameter statistical models such as the memory models just discussed (Gelman et al., 2004, Chapter 5). From the perspective of psychology, hierarchical methods are important and useful since they allow psychologists to model individual differences in performance as well as facilitate the modeling of changes in performance.

4.1. Basic assumptions

Our hierarchical Bayesian analysis requires a few basic assumptions. To motivate these assumptions, consider our clinical data set. In this data set, we should expect variability on at least two qualitatively distinct levels to contribute to the observed memory data. At an individual level, each patient should be expected to have differing memory ability; for example, regardless of any impairment due to ADRD, some patients simply have better memory ability than do others. At a group level, patients with a given FAST stage classification are expected, on average, to differ in their memory ability from patients with a different FAST stage classification. For example, memory performance (e.g., as measured by the parameters of the memory models) should decrease as severity of dementia (e.g., as measured by the FAST) increases.

To formalize these intuitions, we first assume that FAST stage groups are modeled as Gaussian distributions over the possible values of the memory strength parameters (viz., discriminability d', primacy α, and recency β), with the means and variances of these distributions differing both across parameters and FAST stage classifications. Our second basic assumption is that the unique memory strength parameter values for each patient are sampled from the appropriate group distributions that are determined by each patient's FAST stage classification.

One additional assumption is needed in order to implement our hierarchical Bayesian analysis. Each patient completes only one recognition task. However, each patient completes four recall tasks during the dementia assessment, and it is highly unlikely that, for any given patient, performance on these four tasks is independent. Consequently, some assumption that explains how patient performance is related on these tasks is required. Rather than developing a psychological model that ties these tasks together, we performed a simple statistical analysis based on independent implementations of the general model just described. Briefly, the means of the primacy α and recency β group distributions were either increased or decreased, relative to the baseline levels for the first recall task, by an amount that stayed constant across the six FAST stages. A complete explanation of the statistical analysis underlying this model is given in the Appendix. We return to the issue of using statistical versus psychological models to account for this change in the Discussion.

4.2. Graphical model

Although the memory models discussed in the previous section are relatively simple with respect to the psychological assumptions they embody, a full hierarchical analysis with these models greatly increases the complexity of the problem. Thus, it is helpful to find an efficient representations of these models as an aid to both communication and statistical inference. Fortunately, the fields of statistics and machine learning have developed a language that is suited to this task. Graphical models (e.g., Jordan, 2004) provide diagrammatic representations of statistical models in which the nodes of a graph correspond to random variables, and the edges between these nodes correspond to the various independence assumptions of the statistical model the graph represents, with children independent of all other nodes given their parents. We adopt the notation used in recent tutorials on graphical models aimed at psychologists (Lee, 2008; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Square nodes represent discrete variables and circular nodes represent continuous variables. Shaded nodes represent observed quantities and unshaded nodes represent unobserved quantities; in situations where the data are "partially observed" (i.e., when there are missing data), a lighter shade is used to indicate the variable that is partially observed. Stochastic variables are represented by nodes with a single border and deterministic nodes are represented with double borders. Finally, independent replications of portions of the graph structure are enclosed within rectangles, which are referred to as *plates*.

A graphical model representation for our hierarchical Bayesian analysis is shown in Fig. 3. The left-hand side of this graphical model is a representation of the two-factor model for the recall data, and the right-hand side of the model is a representation of the SDT model for the recognition data, which we will now describe in turn.

4.2.1. Two-factor recall model

At the level of FAST stage groups, the means of the primacy α and recency β parameter distributions for stage i on recall task $t = 1$ are given independent, non-informative uniform prior distributions over all possible parameter values, with

$$\mu^{(1)}_{\alpha,i}, \mu^{(1)}_{\beta,i} \sim \text{Uniform}(0, 1).$$

Similarly, the standard deviations of the primacy α and recency β parameter distributions for stage i, which remain constant across the four recall tasks, are given independent, non-informative uniform prior distributions, with

$$\sigma_{\alpha,i}, \sigma_{\beta,i} \sim \text{Uniform}(0, 1).$$

For each subsequent recall task $t \in \{2, 3, 4\}$, the mean of the primacy α parameter distribution for stage i on the first immediate recall task t is given by

$$\mu^{(t)}_{\alpha,i} = \mu^{(1)}_{\alpha,i} + \delta_{\alpha},$$

and the mean of the recency β parameter distribution for stage i on the recall task t is given by

$$\mu^{(t)}_{\beta,i} = \mu^{(1)}_{\beta,i} + \delta^{(t)}_{\beta}.$$

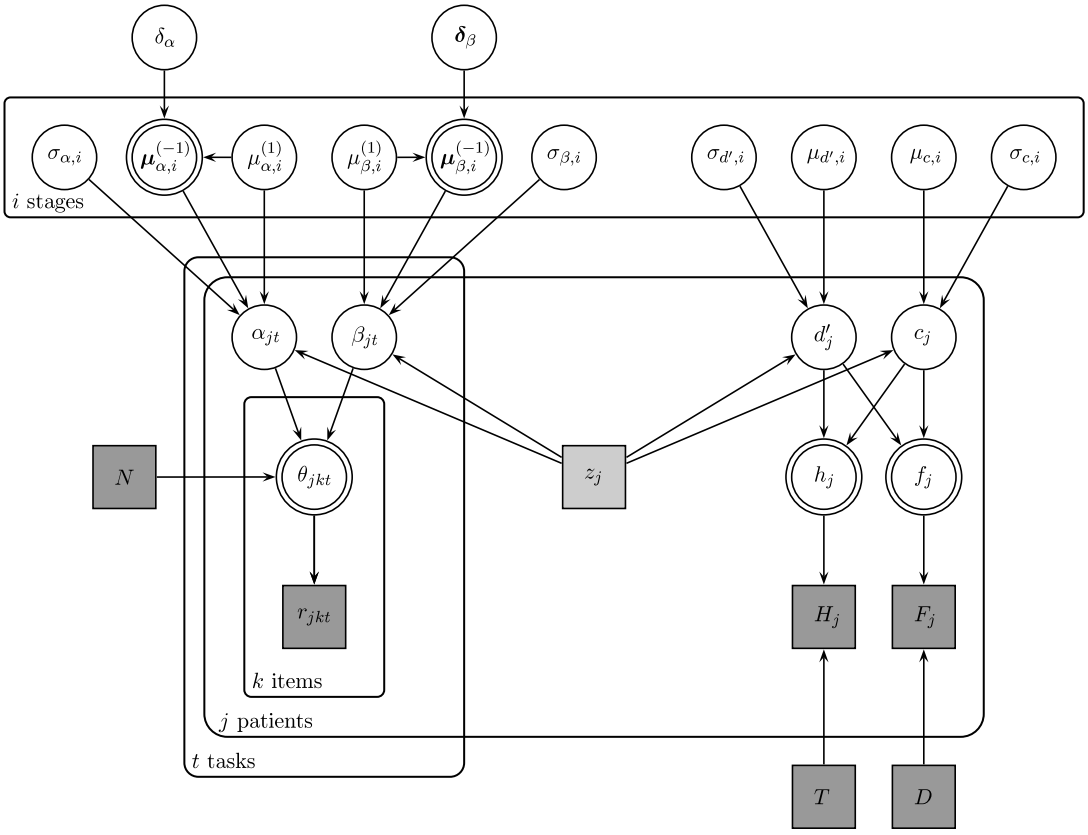

Fig. 3. Graphical model representation for our hierarchical Bayesian analysis.

In the graphical model shown in Fig. 3, these means are combined into deterministic nodes which correspond to the subsequent primacy means vector

$$\boldsymbol{\mu}^{(-1)}_{\alpha,i} = (\mu^{(2)}_{\alpha,i}, \mu^{(3)}_{\alpha,i}, \mu^{(4)}_{\alpha,i})$$

and the subsequent recency means vector

$$\boldsymbol{\mu}^{(-1)}_{\beta,i} = (\mu^{(2)}_{\beta,i}, \mu^{(3)}_{\beta,i}, \mu^{(4)}_{\beta,i}),$$

where the superscript "−1" is intended to indicate that the vector contains the means of "all recall tasks *except* recall task 1". Since each of the six FAST stage groups is assumed to have its own primacy and recency parameter distributions, all of these nodes are enclosed within the plate indexed by $i \in \{1, \ldots, 6\}$.

Based on the statistical analysis discussed in the Appendix, a single primacy change parameter applies across the subsequent recall tasks $t \in \{2, 3, 4\}$ and is given a non-informative uniform prior distribution, with

$$\delta_\alpha \sim \text{Uniform}(-1, 1).$$

Similarly, the recency change parameter each subsequent task $t \in \{2, 3, 4\}$ is also given a non-informative uniform prior, with

$$\delta^{(t)}_\beta \sim \text{Uniform}(-1, 1).$$

In the graphical model, these parameters are combined in the vector

$$\boldsymbol{\delta}_\beta = (\delta^{(2)}_\beta, \delta^{(3)}_\beta, \delta^{(4)}_\beta).$$

At the level of individual patients, the primacy α_{jt} and recency β_{jt} parameters for patient j on the recall task t are drawn from the appropriate distributions at the group level determined by their FAST stage classification, which is given by the categorical variable $z_j \in \{1, \ldots, 6\}$. Specifically, a patient's primacy and recency parameters are drawn from truncated Gaussian distributions with appropriate mean and standard deviation, with truncation below 0 and above 1 since, formally, these parameters represent probabilities of recall. More formally, the primacy parameter

$$\alpha_{jt} \sim \text{Gaussian}(\mu^{(t)}_{\alpha,z_j}, \sigma_{\alpha,z_j}) \times \mathbb{I}_{[0,1]}$$

and the recency parameter

$$\beta_{jt} \sim \text{Gaussian}(\mu^{(t)}_{\beta,z_j}, \sigma_{\beta,z_j}) \times \mathbb{I}_{[0,1]},$$

where $\mathbb{I}_{[0,1]}$ is the indicator function on the unit interval $[0, 1]$. For reasons that will be explained shortly, we withheld the known FAST stage classifications of a small portion of the patients from each FAST stage. The prior on the stage indicator for these patients is

$$z_j \sim \text{Categorical}\left(\frac{1}{6}, \ldots, \frac{1}{6}\right).$$

Since each patient is assumed to have unique parameter values on each recall task, these nodes are enclosed within the plates indexed by $j \in \{1, \ldots, 525\}$ patients and $t \in \{1, \ldots, 4\}$ tasks.

These primacy and recency parameters then deterministically combine to give the probability that patient j recalls word k on the recall task t as

$$\theta_{jkt} = 1 - (1 - \alpha^k_{jt})(1 - \beta^{N-k+1}_{jt}),$$

where $N = 10$ is the number of words on the study list. Given this recall probability, patient j either recalls or fails to recall word k on the recall task t according to the binary random variable

$$r_{jkt} \sim \text{Bernoulli}(\theta_{ijk}),$$

where $r_{jkt} = 1$ indicates that the word was recalled and $r_{jkt} = 0$ indicates that the word was not recalled. Since these nodes are also enclosed within the patient and task plates, as well as the plate indexed by each word (or serial position) $k \in \{1, \ldots, 10\}$ on the study list used for the free recall tasks.

4.2.2. SDT recognition model

At the level of FAST stage groups, the means of the discriminability and bias parameter distributions for stage i are both given non-informative Gaussian priors, with the mean of the discriminability parameter

$$\mu_{d',i} \sim \text{Gaussian}(0,\ 1/\sigma^2 = 2),$$

and the mean of the response bias parameter

$$\mu_{c,i} \sim \text{Gaussian}(0,\ 1/\sigma^2 = 1/2).$$

Following the general advice of Gelman (2006), the standard deviations of the discriminability and response bias parameter distributions for stage i are both given non-informative uniform priors, with

$$\sigma_{d',i},\ \sigma_{c,i} \sim \text{Uniform}(5/100,\ 3).$$

Since each of the six FAST stage groups is assumed to have its own discriminability and response bias parameter distributions, all of these nodes are enclosed within the plate indexed by $i \in \{1, \ldots, 6\}$.

At the individual patient level, each patient j has a unique discriminability and response bias, each of which is sampled from the appropriate group distribution depending on their FAST stage classification z_j, with the discriminability

$$d'_j \sim \text{Gaussian}(\mu_{d',z_j},\ \sigma_{d',z_j}),$$

and the response bias

$$c_j \sim \text{Gaussian}(\mu_{c,z_j},\ \sigma_{c,z_j}).$$

The discriminability and response bias for each patient j is then reparameterized according to SDT into a hit rate h_j and false alarm rate f_j. Finally, based on these hit and false alarm rates and $T = 10$ targets and $D = 10$ distractors, the hit and false alarm counts for patient j follow binomial distributions, with

$$H_j \sim \text{Binomial}(h_j,\ T),$$

for the hits and

$$F_j \sim \text{Binomial}(f_j,\ D),$$

for the false alarms. These nodes are enclosed within the plate indexed by $j \in \{1, \ldots, 525\}$ patients.

4.3. Details of MCMC sampling

The graphical model shown in Fig. 3 was implemented in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), the software which uses a variety of Markov chain Monte Carlo (MCMC) algorithms (e.g., Gamerman & Lopes, 2006) to draw samples from the posterior distributions of the parameters of interest. Our results are based on three chains consisting of 5000 samples collected following a burn-in period of 1000 samples. Convergence of the chains was assessed using the $\hat{R}$ statistic (Brooks & Gelman, 1998).

5. Results

5.1. Posterior predictive distributions

Before examining the posterior distributions for these memory strength parameters, it is important to check that our model is sensible. Many factors contribute to what makes a psychological model sensible, and just which of these factors are emphasized in a given analysis will ultimately depend on both the model itself (e.g., the "level of analysis" at which the model is formulated) and on the context in which the model is applied (e.g., Shiffrin et al., 2008). In our application, it makes sense to focus mainly on our model's descriptive adequacy (i.e., the ability of the model to account for and describe interesting patterns in the observed data). Posterior predictive distributions provide an intuitive and principled graphical approach to assessing the descriptive adequacy of a Bayesian model (e.g., Gelman et al., 2004, pp. 165–172). A posterior prediction corresponds to the future (or, more generally, missing or unobserved) data the model expects, based on the parameter values it has inferred from the observed data, and naturally takes into account uncertainty in these parameter estimates.

5.1.1. Recall data

Fig. 4 shows the group posterior predictive distributions for the recall data. In the figure, the posterior predictions made by the model are shown as squares, with the area of a square proportional to the posterior predictive mass of the data point it represents. It is clear from the figure that, for each block of recall tasks, the group level predictions of the model match the observed serial position curves. Specifically, the model is able to capture the general decrease in performance as the FAST stage increases, and the model also captures the loss of the recency effect in the delayed free recall task (bottom panel of Fig. 4). We note that the fit of the model to the delayed free recall task is noticeably worse than the model fit to the previous three recall tasks. This is quite probably due to the simple nature statistical analysis, described in the Appendix, that ties together patient performance on the four recall tasks. Still, the fit seems good enough for the exploratory purposes of the current application.

5.1.2. Recognition data

Fig. 5 shows the group posterior predictive distributions for the recognition data. It is again clear from the figure that the group level predictions of the model match the observed hit and false alarm counts for the patients and show a general degradation in performance, with fewer hits and more false alarms, as the FAST stage increases.

5.2. Posterior distributions

Given that our model achieves a basic level of descriptive adequacy, it is sensible to examine the posterior distributions for the model parameters shown in Fig. 6. The top panel of this figure shows the joint posterior distribution for the recall primacy α and recency β parameters, and the bottom panel shows the joint posterior distribution for the discriminability d' and bias c parameters. Each column corresponds to one of the six FAST stages. The black dots show 100 samples from the joint posterior distribution of the parameters for the FAST stage of interest and the gray dots show the same for all of the other FAST stages. It is clear from the figure that these parameters change systematically with the FAST stage. In the case of the recall parameters, primacy is affected early in the course of ADRD and continues to degrade with progression through the FAST stages. Recency, on the other hand, seems to be relatively spared early in the course of ADRD, and it is not until late in the course of ADRD (i.e., FAST stage 6) that recency seems to be particularly affected.

In the case of the recognition parameters, the bottom panel of Fig. 6 shows that the six FAST stage groups are well separated in terms of the discriminability parameter. This is intuitive since one would expect memory to be increasingly impaired (e.g., discriminability should decrease) as the severity of ADRD increases. Perhaps less intuitive is the result that recognition bias does not seem to be particularly affected. Instead, bias is relatively conservative for unimpaired patients (i.e., FAST stage 1), and it still is relatively conservative (or at least unbiased) for severely impaired patients (i.e., FAST stage 6). This finding is in contrast to what some researchers have found when investigating the effects of dementia on recognition bias (Budson, Wolk, Chong, & Waring, 2006; Snodgrass & Corwin, 1988).

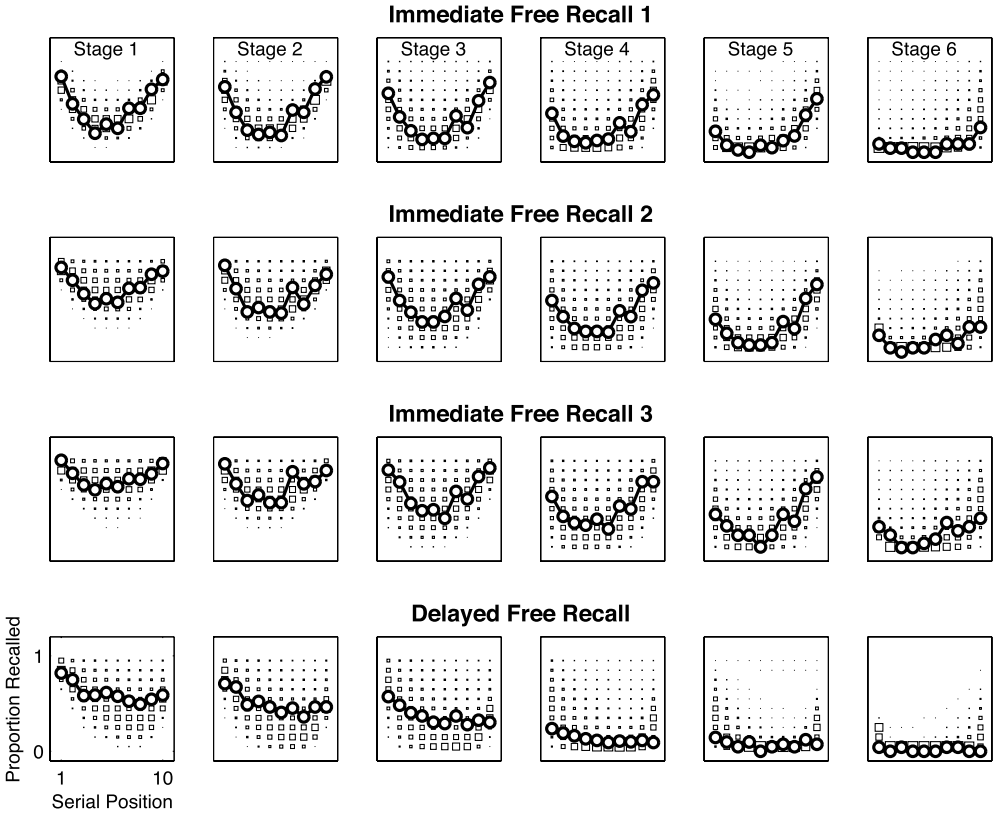

Fig. 4. Posterior predictive distributions for the recall data. The black line connected by open circles shows the observed serial position curve, and squares represent posterior predictions made by the model. The areas of the squares are proportional to the posterior predictive mass.

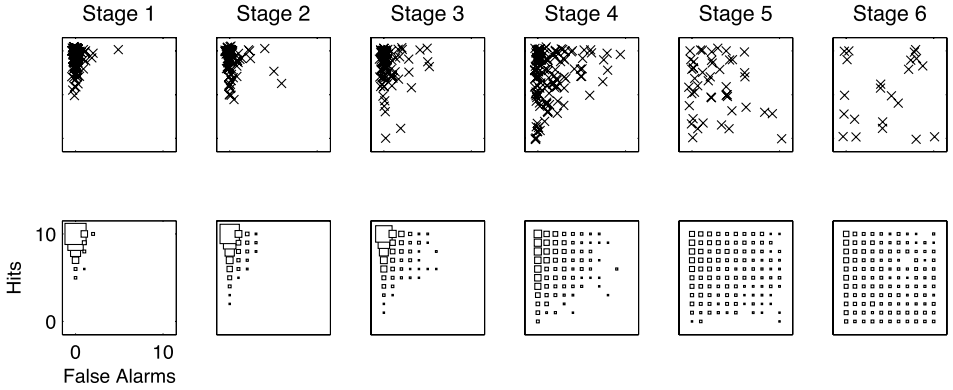

Fig. 5. Posterior predictive distributions for the recognition data. The top panel shows the observed hit and false alarm counts for each patient, and the squares in the bottom panel represent posterior predictions made by the model. The areas of the squares are proportional to the posterior predictive mass. Each column corresponds to one of the six FAST stages.

5.3. FAST stage classification

Given that the task of prediction is naturally accommodated by the Bayesian framework, we decided to test our model on the task of predicting a patient's known FAST stage from their performance on the memory tasks. This was accomplished by using the recognition and recall data from five patients with each stage classification to obtain a posterior distribution over the memory strength parameters. We then found the posterior distribution for the FAST stage variables z_j of those patients whose "true" FAST stage was withheld, and made the predicted stage corresponding to the mode of that distribution. This process was repeated 25 times with a different set of five patients from each stage randomly chosen to be withheld from the model.

The classification results are shown in Fig. 7, which shows how the model's classifications compares to the "true" classifications. Each square corresponds to a truth-prediction pair, and the box size is proportional to the frequency of that pairing. Black boxes on the diagonal are correct classifications. The gray regions in Fig. 7 correspond to the broader classification dividing FAST stages 1 and 2, which essentially represent normal cognitive functioning, from stages 3–6, which represent cognitive impairment with or without dementia (i.e., ADRD). It can be seen that the predictions of the model are generally good, especially at the broader level, but are certainly not perfect. We are aware of the debate in the field of machine learning (e.g., Ng & Jordan, 2002) concerning the relative performance of generative models (e.g., our model) versus discriminative models (e.g., logistic regression) for the task of classification, with the general belief being that discriminative methods are superior. It may well be the case that our method is not optimal at this task. Our main point, however, is that it is straightforward to make predictions for individuals by assuming

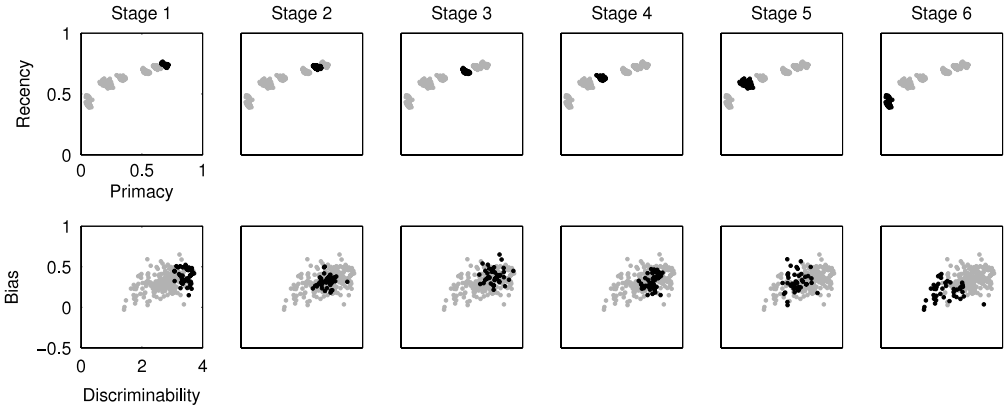

Fig. 6. Joint posterior distributions for the model parameters. The top row shows samples from the joint posterior distribution over the recall parameters, and the bottom row shows samples from the joint posterior distribution over the recognition parameters. Each column corresponds to one of the six FAST stages, with the set of samples from the FAST stage represented by the column shown in black.

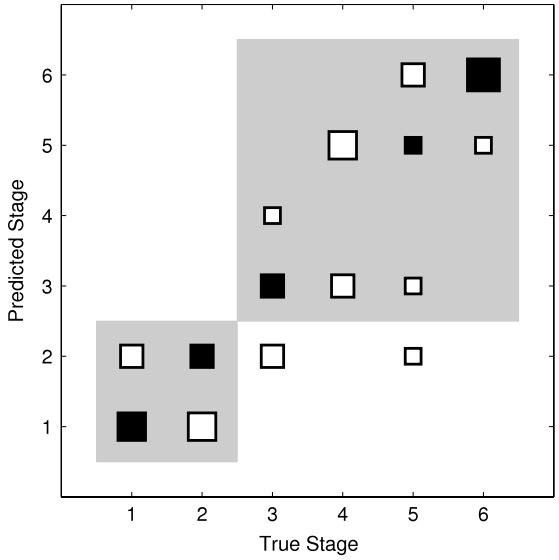

Fig. 7. FAST stage predictions. Black squares represent correct classifications and white squares represent incorrect classifications. The size of each square is proportional to the frequency of that classification. The small gray region represents the broad classification of cognitively normal and the large gray region represents the broad classification of ADRD.

hierarchical individual differences, and that these predictions are informed by the different characteristics observed in Fig. 6.[2]

5.4. Parameter correlations

An interesting analysis is shown in Fig. 8, relating the d' measure of "memory strength" in recognition to the primacy α and recency β "memory strength" parameters in the recall model. It is clear that these parameters covary systematically, consistent with the notion that these parameters tap some common basic property of human memory. It is an open problem to construct some theory that integrates these two processes, which has long been a goal of cognitive and mathematical psychologists interested in a general theory of human memory and could also serve as the basis for better methods for assessing ADRD patient performance in clinical settings. The only point we wish to demonstrate here is that, due to the nature of clinical assessments, the resulting data sets may provide excellent settings in which to develop such a memory theory.

6. Discussion

We feel that our results demonstrate some of the advantages that the combination of cognitive models and hierarchical Bayesian methods have for the understanding of memory impairments in ADRD patients. Clinicians interested in the early detection of ADRD and related concerns as well as psychologists interested in the basic processes of human memory and their dysfunction both should have reasons for adopting this methodology in their research.

As discussed in the introduction, the screening and assessment of ADRD relies on the use of psychological tasks. Using psychological models to analyze the data from memory tasks can aid in the understanding of the results in ways that the ad hoc methods (e.g., counts of correct recognition choices) cannot, and a more thorough understanding of these results should ultimately lead to modifications of the existing psychological tasks used in ADRD assessments that are differentially sensitive to different forms and severities of dementia.

A methodological issue that can be examined with the hierarchical methodology relates to the words used as stimuli for the assessment tasks. The specific subset of words used as stimuli were selected from the CERAD word list with the goal of minimizing the semantic associations between these words. However, even the specific subset of words selected from the CERAD word list are unlikely to be free from these associations. The addition of these "item effects" to our model (cf. "participant effects" included in our model) is naturally accomplished using the hierarchical Bayesian methodology, thus allowing for more accurate estimation of the model's parameters (e.g., Rouder & Lu, 2005). For example, Pratte and Rouder (2011) use this more sophisticated approach that includes distributions over both individuals and items in an application to recognition memory data.

Another area of research that may be worth exploring with the hierarchical methodology adopted here relates to the FAST stages themselves. In this paper, we have assumed that the FAST stages as classified by the clinician represent the "true" level of severity of a patient's ADRD impairment. However, it is possible that the clinician's FAST classification of a patient is not justified by the patient's assessment performance, and it is certainly the case that the discretization of a more or less continuous degradation in memory performance by the FAST is a simplification of a more complicated reality. Thus, issues such as the consistency of a clinician's classifications and the justification of a FAST stage in terms of a patient's memory performance could potentially be usefully evaluated within the modeling framework used here.

[2] Regardless of the optimality of generative versus discriminative classifiers for the task of classification, an anonymous reviewer raised the possibility that part of the reason for the misclassifications could be due to errors on the part of the clinicians making the FAST classifications. This seems reasonable, and is an issue worth further scrutiny.

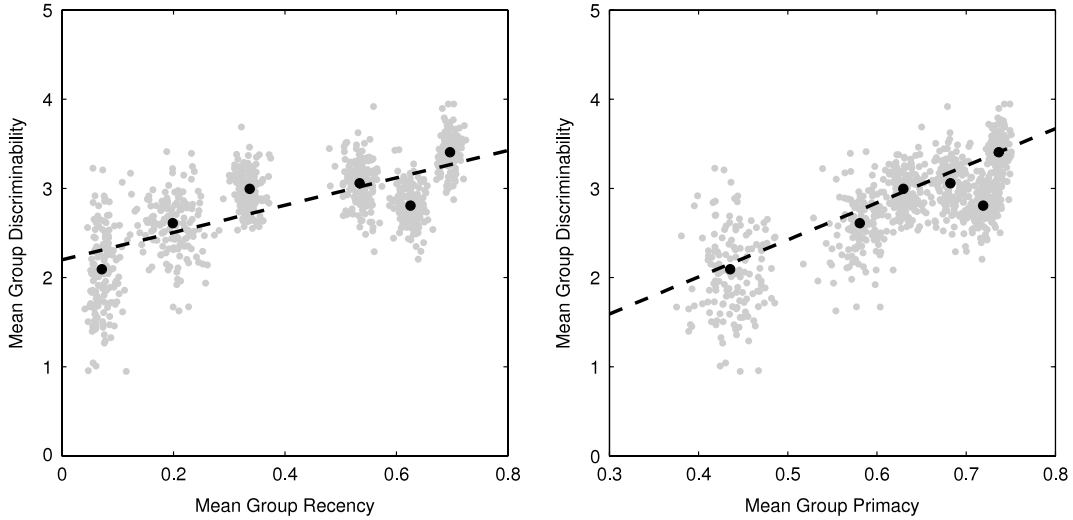

Fig. 8. Correlations between the d' and α and β memory strength parameters. Gray dots represent posterior samples, and black dots represent the means of the samples.

There are several reasons for this work to be of interest to cognitive and mathematical psychologists. In contrast to the majority of work in these disciplines, clinical assessments often result in large data sets such as the one analyzed here. Large data sets such as these can in principle be used to answer questions that may not be possible to answer with the small data sets from undergraduate students typically used in modeling studies. In addition, by forcing psychologists to come to terms with individual differences and systematic changes in memory performance across tasks by asking good (and sometimes new) theoretical questions, clinical data sets such as the one analyzed in this paper force the psychologist to make useful modifications to existing models. This process increases the realism and significance of the models.

Obviously, the measurement models applied in this paper are much more limited in terms of explanatory power than are the process models currently preferred in theoretical studies of human memory. Although, as advocated throughout this paper, we feel that measurement models provide an attractive compromise between theoretical insight and practical utility, we also feel that the application of the more complex process models should ultimately be pursued. For example, as mentioned above, there are different patterns of decay for the primacy and recency parameters of the two-factor recall model. It is easy to see how a more complex recall model (e.g., a two-store model with more detailed processing assumptions) could provide more detailed insight on the nature and importance of this observation. However, this interplay between measurement and process models, for clinical applications and basic research, is an issue that awaits further study.

Ultimately, as we have tried to show throughout this paper, any distinction between developments that are exclusively of either clinical or cognitive relevance is an artificial one. Progress in modeling the basic processes of human memory will naturally inform the practice of assessing when and how these processes fail in patients with ADRD, and progress in clinical screenings (and the large data sets they produce) will naturally lead to the revision of existing cognitive models and to the development of new models that account for previously unmodeled phenomena. In any case, we feel that the hierarchical Bayesian methodology will play a key role in such progress, and feel that the research presented in this paper demonstrates its potential.

Acknowledgments

This research was supported by award NIRG-08-90460 from the Alzheimer's Association. We wish to thank Bill Batchelder, Mark Steyvers, and three anonymous reviewers for their helpful comments which improved the content and presentation of this paper. We also wish to thank Dr. Douglas Trenkle, D.O. for access to portions of the Hancock County Aging Project data set.

Appendix

In this appendix, we discuss the development and details of the statistical model used to tie a patient's performance across the four free recall tasks. As motivated below, this purely statistical model for the change in performance across tasks is based on statistical rather than psychological considerations. However, a psychological theory that explains this change is ultimately to be desired.

Our analysis is based on independently running the same hierarchical SDT model and estimating the means of the primacy and recency parameters from data from each of the four recall tasks independently. The means of the primacy and recency parameters for each of the FAST stage groups were independently estimated from data from each of the four recall tasks. This process is represented by the graphical model in Fig. 9. All results are based

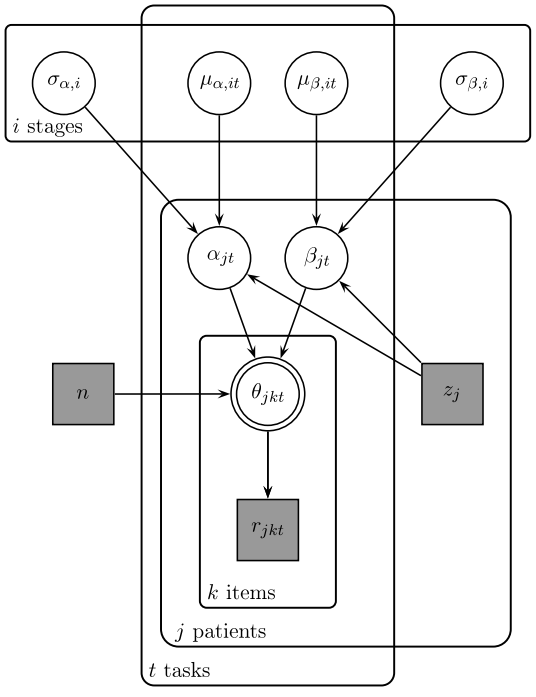

Fig. 9. Graphical model representation for our hierarchical Bayesian analysis.

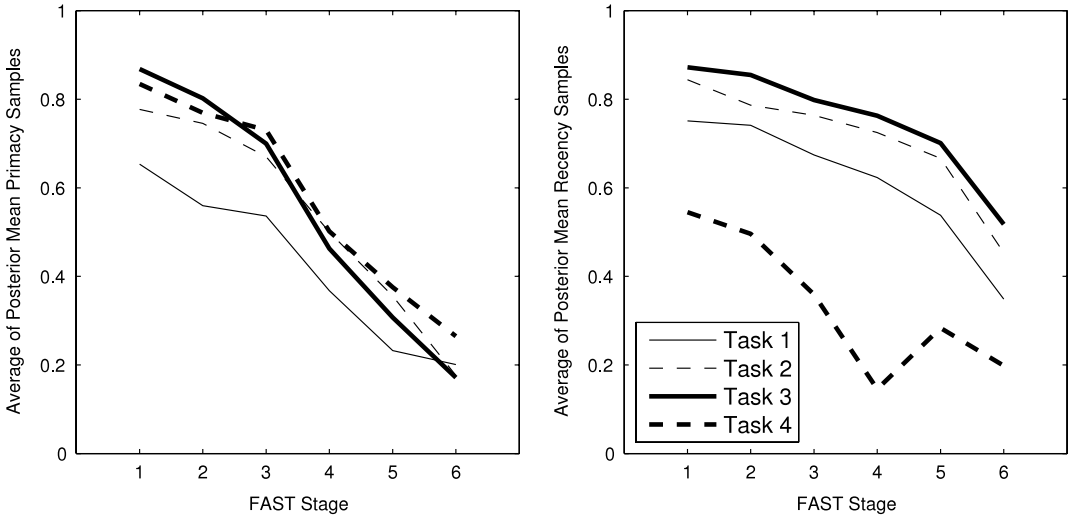

Fig. 10. Averages of the posterior samples of the means of the recall parameter distributions across the four recall tasks.

on three chains consisting of 1000 samples collected following a burn-in period of 100 samples, and convergence of the chains was assessed using the $\hat{R}$ statistic.

Fig. 10 shows, for each of the six FAST stages, plots of the mean of the posterior samples of the mean primacy and recency parameters. Based on the plot of the mean of the primacy parameter in the left panel of Fig. 10, we decided that is reasonable to approximate the change in the mean primacy strength between the tasks as a simple increase in strength δ_α that stays constant across each of the four tasks and each of the six FAST stages, relative to the baseline provided by the first recall task.

The right panel of Fig. 10 shows the plot for the mean of the recency parameter. Based on this plot, we feel that it is reasonable to model the change in the mean of the recency parameter as an increase in strength $\delta_\beta^{(2)}$ on the second recall task, a larger increase $\delta_\beta^{(3)}$ on the third recall task, and a decrease $\delta_\beta^{(4)}$ on the forth recall task.

Despite the relative simplicity and admittedly ad hoc nature of this change model as a way of unifying a patient's performance on the four recall tasks, the model works reasonably well in practice, as demonstrated by the posterior predictives in Figs. 4 and 5. However, as we mentioned above, an improved statistical model or a model motivated by more psychological concerns is obviously desirable.

References

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Budson, A. E., Wolk, D. A., Chong, H., & Waring, J. D. (2006). Episodic memory in Alzheimers disease: Separating response bias from discrimination. *Neuropsychologia*, *44*, 2222–2232.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Henson, R. N. A. (1998). Short-term memory for serial order: The Start-End Model. *Cognitive Psychology*, *36*, 73–137.

Hodges, J. R. (2000). In E. Tulving, & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 441–464). New York: Oxford University Press.

Hodges, J. R. (2006). Alzheimer's centennial legacy: Origins, landmarks and the current status of knowledge concerning cognitive aspects. *Brain*, *129*, 2811–2822.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Locascio, J. J., Growdon, J. H., & Corkin, S. (1995). Cognitive test performance in detecting, staging, and tracking Alzheimers disease. *Archives of Neurology*, *52*, 1087–1099.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Morris, J. C., Mohs, R. C., Rogers, H., et al. (1988). CERAD: clinical and neuropsychological assessment of alzheimer's disease. *Psychopharmacology Bulletin*, *4*, 641–652.

Neufeld, R. W. J. (Ed.) (2007). Advances in clinical cognitive science: Formal modeling of processes and symptoms. Washington, DC: APA Books.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems (NIPS). 14*.

Norman, K. A., Detre, G. J., & Polyn, S. M. (2008). Computational models of episodic memory. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*. New York: Cambridge University Press.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*(1), 36–46.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*, 323–341.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Reisberg, B. (1988). Functional assessment staging (FAST). *Psychopharmacology Bulletin*, *24*, 653–659.

Rouder, J. N., & Lu, J. (2005). An introduction to bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*, 1–25.

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.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.