A Multinomial Modeling Analysis of Memory Deficits in Alzheimer’s Disease and Vascular Dementia

William H. Batchelder, Jamie Chosak-Reiter, W. Rodman Shankle, and Malcolm B. Dick

Departments of Cognitive Sciences, Neurology and Computer Science, and Neurology, Alzheimer’s Disease Research Center, University of California, Irvine.

Data from the immediate recall task of the Consortium to Establish a Registry for Alzheimer’s Disease neuropsychological test battery were disaggregated into nine subject groups and analyzed with traditional statistics as well as with a general processing tree (GPT) model of free recall. The groups represented four levels of severity of Alzheimer’s and vascular dementia, as well as a ninth group of healthy elderly controls. It was demonstrated that the patterns of success and failure of recall to individual items across successive trials contained much more information than the marginal trial-to-trial performance scores traditionally used in scoring the test. The GPT model analyzed recall performance in terms of three levels of item storage: unstored, intermediate, and long-term. Associated with the intermediate and long-term storage levels were respective retrieval parameters. Statistical methods enable one to estimate the parameters for each group, and the analyses revealed group differences in long-term storage that were not evident in a statistical analysis of the marginal trial-to-trial performance scores.

It is widely known that patients diagnosed with Alzheimer’s disease (AD) are impaired in explicit or declarative memory tasks, such as recall and recognition (Hodges & Patterson, 1995; Welsh, Butters, Hughes, Mohs, & Heyman, 1991, 1992). Understanding the nature of these impairments in AD remains a challenge for dementia researchers, mainly because deficiencies in different subsets of underlying or latent cognitive processes can account for the same levels of reduced performance in a memory task. For example, in free recall performance, any number of component memory processes may be operating (Nebes, 1992), such as encoding items, maintaining these items in short-term or temporary memory, forgetting, longer-term learning, as well as primacy and recency effects. Unfortunately, few traditional psychometric techniques have been developed to assess latent cognitive processes. Traditional ad hoc statistical methods (e.g., log-linear or analysis of variance [ANOVA]) are useful for detecting statistical differences between populations or experimental conditions based on observed performance scores, but they are not designed to assess and compare the various contributions of underlying cognitive processes to these performance scores because they do not model the subprocesses. There have been a few attempts at measuring some of the cognitive processes underlying memory performance by operationally defining performance scores on a free recall and recognition task, for example, Geffen, Moar, O’Hanlon, Clark, and Geffen (1990); Mitrushina and Satz (1989); and Mitrushina, Satz, Chervinsky, and D’Ella (1991). A more detailed discussion of these studies will be taken up at a later stage.

In addition to the problem of understanding the specific nature of the memory impairment in AD, distinguishing between the early stages of AD and vascular dementia (VD) based on neuropsychological tests remains a challenge for investigators (Almkvist, Backman, Basun, & Wahlund, 1993; Carlesimo, Fadda, Bonci, & Caltagirone, 1993; Villardita, 1993). Lower performance scores on explicit memory tasks are observed with increasing severity in AD and between early AD and healthy controls (Fillenbaum, Wilkinson, Welsh, & Mohs, 1994), but understanding which component memory processes contribute to the declining performance is difficult, if not impossible, without an explicit model of the task.

In an effort to differentiate between early AD and controls, memory performance has been analyzed using traditional statistical methods (Hodges & Patterson, 1995; Welsh et al., 1991, 1992). Although these researchers were successful in finding differences with some of the measures available to them, other tasks did not prove to be as useful. For example, Welsh et al. (1991, 1992) found that mean scores for the immediate recall task of the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) neuropsychological test battery did not differentiate between mild AD and healthy control subjects, whereas the delayed recall task did differentiate these two groups. Further, they suggested that the delayed recall task was not useful in differentiating the different stages of AD.

In this article, the investigators present a new statistical method that analyzes the specific pattern of item recall over three study-test trials as well as the mean performance scores and estimates the differential contributions to performance of several underlying cognitive processes. Evaluating data in this way allows more information to be obtained, which in turn increases the possibility of finding differences between subject populations when they exist, as well as providing greater understanding of the nature of the memory impairment in dementia.

The method used is to construct a general processing tree (GPT) model of the recall process. GPT modeling is a form of multinomial modeling of categorical data, described in Hu and Batchelder (1994), Batchelder and Riefer (1986, 1990), and Riefer and Batchelder (1988). It is a relatively simple, statistically based technique for estimating the probabilities of hypothetical cognitive processes that underlie...
performance. To use GPT modeling, researchers need to identify the underlying theoretical cognitive processes to be explored and determine how they influence task performance, in the form of category frequencies. A GPT model postulates parameters representing the probabilities of various cognitive processes, and the model is represented as a processing tree, where each branch in the tree is a possible sequence of processing steps leading to a particular response category. If the model is constructed according to the GPT modeling requirements in Hu and Batchelder (1994), then computational software (Hu, 1991) exists that obtains parameter estimates and confidence intervals for these cognitive processes. Also, the software can conduct statistical goodness-of-fit tests of the model as well as hypothesis tests that compare the parameters for different groups of subjects.

The main advantage of this type of methodology is that information can be obtained about underlying cognitive processes based on an explicit model of the observed behavior, as opposed to applying general purpose methods for data analysis that lack explicit substantive assumptions. The GPT methodology is quite flexible with regard to its application. Within the psychological realm, this type of modeling is becoming increasingly popular, and it has been used to examine a variety of different phenomena including storage and retrieval of clusters, source monitoring, proactive and retroactive inhibition, process dissociation, and the recognition-failure paradigm (see, for example, Riefer & Batchelder, 1998; Batchelder & Riefer, 1986, 1990). In all of these applications, as well as many others, GPT models express category probabilities as nonlinear functions of the underlying parameters. These expressions are motivated by an explicit, substantive psychological analysis of the cognitive processes underlying the particular data paradigm. Thus, the GPT approach contrasts with loglinear and logit modeling of categorical data that are general purpose statistical models with linearity assumptions at some level.

In the present study, a GPT model of the free recall component of the CERAD word list memory task was constructed to focus on differences in the immediate and delayed free recall ability of patients with AD, VD, and age-matched healthy controls. The model is capable of determining whether the free recall impairment in AD and VD is due to various combinations of deficits in temporary storage, retrieval from temporary storage, more "permanent" (in the immediate recall context) storage, or retrieval from more permanent storage.

This article is organized into four main sections. Following the introduction, the CERAD immediate recall task is described, and two analyses of the data are presented. The first is a traditional analysis of the CERAD immediate recall data for AD, VD, and elderly controls, and this is followed by an analysis of the same data using the GPT model. Finally, the conclusion stresses the value of the model-based approach.

**METHOD**

**Participants**

The clinical sample consisted of 381 patients enrolled in the University of California, Irvine, Alzheimer’s Disease Research Center. Only English-speaking subjects who completed the full assessment were included. Diagnoses were based on standard physical and neurological examinations performed by a faculty neurologist and a comprehensive 2-hour battery of cognitive tests, including the CERAD battery, administered by a licensed neuropsychologist. In addition, all subjects received routine laboratory blood analyses, electrocardiogram, chest X-ray, social and family interviews, and neuroimaging (magnetic resonance imaging with or without SPECT).

AD subjects satisfied National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s Disease and Related Disorders Association criteria for “probable” or “possible” AD (McKhan et al., 1984). Clinical diagnoses of probable and possible AD are equally valid with regard to postmortem neuropathological confirmation of AD (Edland, Morris, Heyman, & van Belle, 1992) and approximately 90% of probable AD diagnoses are confirmed at autopsy (Kazee et al., 1993). Of our sample, 121 subjects were classified as probable AD and 96 as possible AD. VD subjects satisfied valid criteria of the State of California, Alzheimer’s Disease Diagnostic and Treatment Centers for probable or possible VD (Chui et al., 1992; Corey-Bloom, Galasko, & Thal, 1992) with evidence of cerebral infarction as documented by T1 weighted MRI brain imaging. The probable and possible VD subjects were combined into one group. Patients diagnosed with both AD and VD were excluded from the sample, and a small number (about 4%) of the AD diagnoses were changed to a diagnosis of VD on later visits.

Dementia patients were divided into four severity groups, according to scores on the Mini-Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 1975). The MMSE is composed of 11 questions designed to tap into general cognitive functioning, including attention and recall; scores can range from 0 to 30. The severity grouping was motivated by the severity level cut-off scores from previous studies (Almkvist et al., 1993; Welsh et al., 1991). Patients in the severe groups were defined by MMSE scores between 0 and 14, the moderate groups had scores between 15 and 20, the mild groups had scores between 21 and 25, and the very mild groups had scores between 26 and 30. Patients in the very mild group were distinguished from controls based on mild impairment in other cognitive tasks as well as functional impairment. Breakdowns of subjects and demographic information are provided in Table 1. The breakdowns include the clinical dementia rating (CDR) scores for each subject group. These CDR scores (ranging from 0 = no dementia to 3 = severe dementia) reflect capacity in a range of tests including cognitive skills and daily living activities. A chi-square test of independence between CDR scores and MMSE-derived severity groups combined for AD and VD was rejected, \( \chi^2(12, n = 381) = 347.62, p < .001 \), indicating that there is a strong positive relationship between the two measures of dementia severity. Education level was compared between subject groups using a one-way ANOVA, revealing a significant difference \( F(8,486) = 11.09, p < .05 \).

In an attempt to determine whether education level influenced performance on the recall task, a Pearson’s \( r \) correlation was conducted between education level and total number correct, resulting in an \( r = .27 \). Since \( r^2 = .07 \), this
ANOVA (Group X Trial) for the immediate recall trials with recalled for each group, for each of the three trials as well as the performance on the two delayed test trials. A two-way analysis indicates that education level was not a major determinant of performance scores. Further, because the correlation between MMSE and education was only \( r = .27 \), the MMSE scores were more reflective of dementia severity than education.

**Task Description**

A five-trial free recall task was administered to subjects as part of the CERAD neuropsychological test battery. On each of the first three trials, 10 words (e.g., butter, arm, shore) were visually presented to the subjects to study at a rate of one word every 2 seconds. The subject’s task was to read aloud the words and try to retain them for a later recall test. Each study trial uses a different presentation order of the same 10 words, which was fixed for all participants. Immediately following each study trial was a test of free recall, in which the subject was asked to recall verbally the list words in any order. In addition, there were two delayed recall test trials, at approximately 5 and 30 minutes, respectively, following the third trial. These delayed tests followed unrelated cognitive tests but did not follow additional study trials. A few patients did not participate in the 30-minute delayed trial because it was not added to the testing procedure until recently; therefore, data are not available for the earlier patients in the study.

**RESULTS**

**Learning Curve Analysis**

All statistical tests were conducted at the \( p < .05 \) significance level. Figure 1 illustrates the mean number of items recalled for each group, for each of the three trials as well as the performance on the two delayed test trials. A two-way ANOVA (Group \( \times \) Trial) for the immediate recall trials with individual subjects as replicates revealed a significant interaction, \( F(16,1120) = 10.01 \) and significant main effects for both Group, \( F(8,560) = 241.06 \), and for Trial, \( F(2,1120) = 414.79 \), indicating that recall performance increased over the immediate test trials and decreased with increasing severity. Post hoc comparisons were conducted on the control, very mild, and mild groups using Tukey tests, all at \( p < .05 \). Significant differences were found on all five trials between the controls and very mild AD and between controls and very mild VD. On the immediate recall trials, the very mild and mild AD groups differed, as well as the very mild and mild VD groups. Only on the delayed trials were significant differences found between very mild AD and VD patients.

The learning curve analysis presented in Figure 1 is based on summary performance statistics that represent a considerable reduction of the data available on the first three trials of the CERAD task. Table 2 depicts these data for a hypothetical subject, giving complete chronology of the errors and successes to each item over the three immediate and two delayed trials, where 1 denotes recalled and 0 failure to recall. The representation shows that there are 50 separate binary scores potentially showing a rich performance pattern on the test. The five column scores provide aggregate performance scores for the subject, and it is these scores that contribute to the learning curve analysis reported in Figure 1. Thus, the complete data underlying the learning curve analyses may have additional information relevant to the cognitive processes underlying recall. The investigators will show that it is possible to use these item performance patterns to estimate probabilities of various cognitive subprocesses underlying the memory processes for individual items.

**A General Processing Tree Analysis**

The model. — The investigators’ strategy in modeling the CERAD data is to construct a model for the individual item performance in free recall that can be applied to data from the first three study-test trials of the task. After the model is constructed, an additional goal is to find a way to use the model to account for the two delayed test trials. This strategy was dictated by the fact that the available data

---

**Table 1. Demographic Data for AD, VD, and Elderly Control Subjects**

<table>
<thead>
<tr>
<th>Age at Test</th>
<th>Education</th>
<th>Males</th>
<th>MMSE</th>
<th>CDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AD (N)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild (21)</td>
<td>69.2 (7.2)</td>
<td>14.8 (3.0)</td>
<td>11 (52)</td>
<td>27.4 (1.2)</td>
</tr>
<tr>
<td>Mild (59)</td>
<td>72.6 (8.5)</td>
<td>13.8 (3.1)</td>
<td>28 (48)</td>
<td>22.8 (1.4)</td>
</tr>
<tr>
<td>Moderate (86)</td>
<td>74.3 (9.0)</td>
<td>13.4 (3.3)</td>
<td>38 (44)</td>
<td>17.8 (1.7)</td>
</tr>
<tr>
<td>Severe (81)</td>
<td>71.0 (9.3)</td>
<td>13.1 (3.3)</td>
<td>38 (47)</td>
<td>9.3 (3.9)</td>
</tr>
<tr>
<td><strong>VD (N)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild (41)</td>
<td>75.4 (7.9)</td>
<td>15.8 (3.2)</td>
<td>25 (61)</td>
<td>28.1 (1.5)</td>
</tr>
<tr>
<td>Mild (37)</td>
<td>77.1 (8.1)</td>
<td>15.0 (3.6)</td>
<td>9 (24)</td>
<td>22.9 (1.7)</td>
</tr>
<tr>
<td>Moderate (30)</td>
<td>78.6 (6.8)</td>
<td>10.6 (4.3)</td>
<td>9 (30)</td>
<td>17.6 (1.9)</td>
</tr>
<tr>
<td>Severe (26)</td>
<td>79.3 (5.8)</td>
<td>11.0 (4.7)</td>
<td>4 (15)</td>
<td>9.7 (3.8)</td>
</tr>
<tr>
<td>Controls (160)</td>
<td>67.5 (12.2)</td>
<td>14.9 (2.7)</td>
<td>110 (68)</td>
<td>29.1 (1.1)</td>
</tr>
</tbody>
</table>

Notes: AD = Alzheimer’s disease; VD = vascular disease; MMSE = Mini-Mental State Examination; CDR = clinical dementia rating score.
on individual items were coded only for the first three trials, and only performance on the entire list was available for each subject who took the two delayed test trials.

The model employed is adapted from the considerable work on Markov learning models conducted in the 1960s and 1970s (see Greeno & Bjork, 1973; Wickens, 1982, for reviews). It is assumed that on any trial an item can be in one and only one of three states of memory storage, depicted in Figure 2. The unstored state (U) corresponds to a state where the item has not yet been encoded and stored into memory and hence necessarily leads to a failure to recall the item; the intermediate state (I) corresponds to a partial or intermediate state of relative weak storage, where item retrieval is possible but may be difficult and transient; and finally the long-term memory state (L) represents a state where the item storage is stronger and more permanent, and recall is more likely. It is assumed that if an item enters state L, it remains there for the remainder of the task.

Figure 2 represents the transition probabilities among the three possible storage states for an item. The model makes two strong but traditional Markovian assumptions: (1) that knowledge of the state of the item on any trial completely determines the transition probabilities to the new state for the item on the next trial, and (2) that the state a subject is in on any trial completely determines the probability of a correct recall response. The retrieval probabilities in each state are shown below, where retrieval from the unstored state is not possible (probability 0), retrieval from the intermediate state occurs with probability t, and retrieval from the long-term memory state with probability I, that is:

\[ P(\text{correct} \mid \text{state } U) = 0 \]
\[ P(\text{correct} \mid \text{state } I) = t \]
\[ P(\text{correct} \mid \text{state } L) = I, \]

where \(0 \leq t, I \leq 1\). Further, since the intermediate memory state I represents weaker storage than the long-term memory state L, the model requires that \(t \leq I\).

The model has a total of four parameters: two storage parameters, \(r\) and \(a\), and two retrieval parameters, \(t\) and \(I\). The parameter \(r\) represents the probability that some memory
storage occurs on a study trial when an item is in state U, and this results in a transition out of state U in the model. The other storage parameter, \( a \), represents the probability of achieving a strong and more permanent storage given that some storage occurs (or has occurred on an early trial). Thus, the transition from U directly to L occurs with probability \( ra \), whereas the transition from U to I is with probability \( r(1-a) \). Once in state I, further transition to L on any trial is with probability \( a \). The two retrieval parameters, \( t \) and \( l \), represent the probability of a correct item recall from states I and L, respectively. Each of the four parameters represents the probability of a particular cognitive subprocess occurring, so they are restricted to fall in the range \( 0 < a, r, t, l < 1 \).

**Data events/response patterns.** — To analyze the recall data using this Markov model, the data needed to be coded into events that could reflect the consequences of transitions from state to state. The present model incorporates three transitions to reflect the three study-test recall trials. The data from these trials were coded into eight events, or response patterns, for each item, reflecting whether or not the item was recalled over the three trials of the CERAD free recall test, although the model could be applied to an arbitrary number of test trials. If an item was recalled it was coded as a 1, and if it was not recalled it was coded as a 0. For example, the data event 101 for a particular item is interpreted as the subject recalling that item on the first and third trials and not recalling it on the second trial. The equation used can be obtained by writing the first author.

<table>
<thead>
<tr>
<th>Item</th>
<th>Immediate (Study-Test)</th>
<th>Delayed (Test Only)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trial 1</td>
<td>Trial 2</td>
</tr>
<tr>
<td>Butter</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Arm</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Shore</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Letter</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Queen</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cabin</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pole</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ticket</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Grass</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Engine</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

*Note: 1 = item recalled; 0 = item not recalled.*

Figure 2. Trial-to-trial transition probabilities among the unstored (U), intermediate (I), and long-term memory (L) states. Each arrow represents the possible transitions from one state to another for each study trial. The probability for each transition is indicated by the parameters, where \( r \) reflects temporary storage and \( a \) reflects more permanent storage.

Table 2. Immediate and Delayed Recall Data From a Typical Subject

Table 3. Counts for the Eight Data Events/Response Patterns for the Three Immediate Recall Trials

<table>
<thead>
<tr>
<th>Data Event</th>
<th>000</th>
<th>010</th>
<th>001</th>
<th>011</th>
<th>101</th>
<th>111</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild</td>
<td>47</td>
<td>13</td>
<td>15</td>
<td>36</td>
<td>12</td>
<td>31</td>
<td>10</td>
</tr>
<tr>
<td>Mild</td>
<td>218</td>
<td>20</td>
<td>56</td>
<td>26</td>
<td>87</td>
<td>30</td>
<td>61</td>
</tr>
<tr>
<td>Severe</td>
<td>620</td>
<td>20</td>
<td>53</td>
<td>10</td>
<td>18</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>VD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild</td>
<td>87</td>
<td>11</td>
<td>28</td>
<td>67</td>
<td>29</td>
<td>71</td>
<td>24</td>
</tr>
<tr>
<td>Mild</td>
<td>131</td>
<td>19</td>
<td>33</td>
<td>56</td>
<td>14</td>
<td>57</td>
<td>11</td>
</tr>
<tr>
<td>Severe</td>
<td>158</td>
<td>18</td>
<td>29</td>
<td>37</td>
<td>8</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>Controls</td>
<td>90</td>
<td>49</td>
<td>53</td>
<td>180</td>
<td>97</td>
<td>384</td>
<td>110</td>
</tr>
</tbody>
</table>

*Note: AD = Alzheimer’s disease, VD = vascular disease. Item recalled denoted by 1, item not recalled denoted by 0.*
tended to handle the two delayed test trials by assuming that correct recall on a delayed test is possible only if the item was in state L following the third study trial. This is consistent with the definition of state I as a temporary state of storage. In this case, let \( l_t \) and \( b_t \) be the retrieval probabilities from state L on delayed Trials 4 and 5, respectively. Since retrievability is likely to decline with time, it is substantively reasonable to require \( l_t \leq l \) and \( b_t \leq l \). With these restrictions in mind, it is possible to characterize the model parameters by the vector \((r, a, t, l, a, l)\), subject to the constraints above and of course that probabilities must be between zero and one.

The available data do not include the particular item responses on Trials 4 and 5, but only the number (out of 10) that were correct. Thus, it was not possible to work with 5-tuples of errors and successes as in Table 2. Instead, the model was fit to the 3-tuples in Table 3, and \( l_t \) and \( b_t \) were estimated from the proportion of correct responses on Trials 4 and 5 as presented in Figure 1. More particularly, after parameter estimates were obtained from the first three trials, the probability of being in state L, \( \Pr (L) \), was calculated by adding the probabilities of all the branches leading to a transition to state L on any of the three trials. The probability of a correct retrieval on Trials 4 and 5 is given by \( \Pr (L) \) times the corresponding retrieval probability, \( l_t \) and \( b_t \), respectively. Thus, it is possible to estimate these retrieval parameters by the formula

\[
\hat{l}_i = \frac{\Pr (\text{correct Trial } i)}{\Pr (L)}
\]

for \( i = 4, 5 \).

**Model fit and results.** — Solutions for the parameter estimators \((r, a, t, l)\) were found using the approach described in Hu and Batchelder (1994) that maximizes the likelihood function for the model. In particular, counts from the actual recall data in Table 3 were inserted into the equations for \( r, a, t, \) and \( l \), and a computer program by Hu (1991) and described in Hu and Batchelder (1994) was used to obtain estimates of the parameters and assess the goodness-of-fit of the solution. In total, nine different sets of estimated parameters were obtained, one for each of the nine subject groups in Table 3.

The goodness-of-fit of a GPT model is indexed by the value of the log-likelihood ratio statistic \( G^2 \). Under suitable assumptions (see Riefer & Batchelder, 1988), \( G^2 \) is approximately chi-square distributed with degrees of freedom equal to the number of categories (8 in this case) less the number of parameters plus one (5 in this case). Small values of \( G^2 \) indicate good fits of the model, so one is looking for values of \( G^2 \) that are sufficiently small. From a chi-square table, the \( p = .05 \) critical value for a chi-square with 3 df is 7.81, and this was the target value to infer goodness-of-fit.

In every case but the control and moderate AD groups, it was possible to achieve the target value for \( G^2 \) (3). However, for the control and moderate AD groups, respectively, \( G^2 \) (3, \( N = 1600 \)) = 11.85 and \( G^2 \) (3, \( N = 860 \)) = 11.74. Since there were nine independent tests of goodness-of-fit, it is not too surprising that two of them did not reach the target value. The model is only approximate in any case, as with any simple statistical model, and is technically false. In particular, with 1600 observations in the control group, the power to reject the model is sufficiently large that it is not surprising that a value exceeding the target was achieved. Riefer and Batchelder (1988, 1991) have a detailed discussion about the issues of model fit and model approximation. In any event, by standards established in other published work with GPT models, the degree of fit is quite satisfactory to warrant further analyses of the data with the model.

A problem was encountered in several of the groups when the program did not constrain the retrieval parameters by \( t < l < u, l \). In these cases, there were several combinations of estimates of the parameters that had different but acceptable \( G^2 \) values, and without some outside criterion it was difficult to justify statistically one choice over the other. More technically, in a few of the groups there were local maxima in the likelihood space that all satisfied the goodness-of-fit criteria and that had different combinations of parameter values. However, when we imposed requirements of a global maximum along with the substantive restrictions on the retrieval parameters discussed earlier, unique best fits were obtained and are reported in Table 4. Table 4 also reports the corresponding parameter estimates for the model, and it includes \( \Pr (L) \) as well as estimates for \( l_t \) and \( b_t \).

It is interesting to examine the pattern of parameters in Table 4. Generally, the parameters tended to decrease with increasing severity level, and the control group's parameters were larger than the other groups. There were, however, several exceptions, and in some cases confidence intervals for the estimates were so large as to prevent confident assessment of the trends. (Confidence intervals are obtained using the observed Fisher information matrix; this is an asymptotic method described in detail for GPT models in Hu and Batchelder [1994]. The approximation entailed may not give accurate confidence intervals when the data relevant to a particular parameter are sparse, or contrary to the model assumptions, when there is a lot of individual parameter variability within a group.) In fact, performance in the moderate and severe groups was sufficiently low that little can be said with confidence about comparing their parameters with the other groups. However, it is true that \( \Pr (L) \) showed reliable declines over all severity levels, and the overall trends in the storage parameters, \( r \) and \( a \), showed a reliable decline with increasing severity. The least reliable (largest confidence intervals) were for the parameter \( t \). This probably reflects the fact that the recency items (positions 9 and 10 during study) performed much better on immediate test than the other items, and the current model does not include a process for parameter inhomogeneity reflecting the recency effect.

In the remaining analyses we concentrated on the very mild and mild groups in relationship to each other and the controls because those groups showed sufficient performance levels for analysis. The approach was to select a pair of groups and equate all four of the parameters between the groups. This yields a four-parameter model for the frequency counts in both groups, so the appropriate goodness-of-fit statistic is a \( G^2(10) \), which is asymptotically distributed as a
not equated. The results were quite revealing. Hypotheses about nested models is that a good fit in this case was achieved by equating all but parameter $a$, $G^2(9) = 11.14$, $p > .05$. The conclusion is that the very mild VD group has a higher value of the long-term storage parameter than the very mild AD group. This is also reflected in the Pr (L) values in Table 4, where very mild AD had Pr (L) = .58, and very mild VD had Pr (L) = .80. Further examination of the columns for $l$ and $h$ suggests that very mild VD retrieve from long-term storage better than very mild AD. An examination of the delayed test trial performance in Figure 1 reveals a difference in favor of very mild VD in delayed recall, and the model analyses suggest that this is due to both long-term storage and long-term retrieval.

Comparing the controls with the very mild groups revealed that significant improvements in the $G'$ could be obtained by allowing only the parameter $a$ to vary between groups. As before, it appears that the basic difference between early dementia and the controls is primarily in long-term storage. This is further suggested by the estimates in Table 4, where Pr (L) is higher for controls than either of the very mild groups, although less so for very mild VD. Further, from Table 4, very mild VD, but not very mild AD, may differ from controls by a substantial amount in the retrieval parameters, $t$ and $l$.

Discussion of model results. — In contrast to the graphs presented in Figure 1, the results of the model's analysis not only differentiated between very mild AD and VD on the data from the first three trials, but also provided a possible explanation for the difference in performance between the stages within AD and VD, and between both dementia groups and controls. An examination of Table 4 suggested that the nature of the differences between AD and controls was different from those between VD and controls.

Compared with very mild AD, the controls exhibited a

Table 4. Parameter Estimates, 95% Confidence Intervals (in Parentheses) and Goodness-of-Fit Indices for the Nine Subject Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>$G'$</th>
<th>$r$</th>
<th>$a$</th>
<th>$t$</th>
<th>$l$</th>
<th>Pr (L)</th>
<th>$l_r$</th>
<th>$h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild</td>
<td>3.22</td>
<td>.99</td>
<td>.25</td>
<td>.20</td>
<td>.87</td>
<td>.58</td>
<td>.53</td>
<td>.43</td>
</tr>
<tr>
<td>Mild</td>
<td>2.59</td>
<td>.95</td>
<td>.30</td>
<td>.05</td>
<td>.69</td>
<td>.65</td>
<td>.22</td>
<td>.18</td>
</tr>
<tr>
<td>Moderate</td>
<td>11.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>3.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>1.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>5.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>11.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Parameters $r$ and $a$ reflect temporary and permanent storage processes, respectively, and parameters $t$ and $l$ reflect temporary and permanent retrieval processes, respectively. Also the performance in delayed test trials for VD severe was so low as to preclude estimating $l$ and $h$ in this case. $G'$ is the log-likelihood ratio statistic. Pr (L) = probability of being in state L, AD = Alzheimer's disease, VD = vascular disease.
much higher value of \( a \), the long-term storage parameter, showed approximately equal values for parameters \( t \) and \( l \); and had higher delayed retrieval values for \( l \) and \( b \). This pattern of change in the parameters suggests the earliest cognitive impairment in AD to be storage of new information on a time scale of minutes and delayed retrieval of that information on a time scale of a few minutes. The storage or perception and learning of new information is thought to be a function of anterior hippocampal circuitry among other cortical structures, while the posterior hippocampus is involved in recognition of newly learned information (Roland & Gulyas, 1995), such that dysfunction of the anterior and posterior hippocampus early in AD may explain this pattern of change in the parameters. It should be noted that the decline in the delayed retrieval probabilities \((l_t \) and \( b_t \)) from state \( L \) could be explained by mechanisms operating on delayed storage of items. This is because the calculation of the probability of being in state \( L \), \( Pr(L) \), depends only on the first three trials; thus, it does not measure storage gains or losses (if any) between the third trial and the delayed recall trials.

The present findings also raise the question of how to relate the parameter estimates for the AD patients to two current theories of the memory impairment in AD. The structural theory, described by Martin (1987), states that the semantic memory stores in AD patients are degraded and/or disorganized. In free recall this loss of semantic information would result in the inability to attach meaning to the study items, making the items more difficult to remember for recall. The retrieval hypothesis suggested by Nebes (1989) states that semantic memory is in large part intact in AD patients, but is inaccessible due to retrieval demands. That is, certain tasks that require effortful attempts at retrieval (e.g., free recall) are more difficult for AD patients, but the same patients can demonstrate access to semantic memory stores through tasks such as priming.

In the present study the estimates for the retrieval parameters are high for the moderate and severe AD patients, but they are unstable because there is very little storage occurring. Therefore, no strong conclusions can be made with respect to these two groups. The estimates for the very mild patients appear not to conform to the retrieval hypothesis, because the estimates for the retrieval parameters do not reflect an impairment relative to controls. In addition, the estimate for parameter \( a \) is lower than that for controls, suggesting a storage deficit. Taken together, the findings for the very mild AD patients seem to support the structural hypothesis. However, in the case of the mild AD patients, the retrieval parameter estimates are lower than those for controls, which is consistent with the retrieval hypothesis. Still, the mild AD patients also demonstrated an inferior level for the storage parameter, \( a \), compared with controls, again favoring the structural hypothesis. It is also possible that both hypotheses contribute to the present findings and to the AD episodic memory impairment in general. Further, because the hypotheses are based on converging evidence from several studies, there may be a wide range of severity levels on which the theories are based. Therefore, it may not be practical to compare the findings for very mild and mild groups in the present study with theories based on potentially more heterogeneous and severely impaired groups of patients.

Compared with very mild VD, the controls exhibited higher values for all retrieval parameters, as well as an advantage in long-term storage. Structures mediating immediate and delayed retrieval of new information appear more affected in VD, which is a subcortical white matter disease in this patient sample. As it is known that VD patients show early impairments in frontal lobe function (Sultzer et al., 1995), an impairment of working memory may affect both temporary and more durable retrieval of new information.

For the very mild groups, VD patients appeared to have better storage than AD patients, measured by parameter \( a \). For very mild dementia, the better long-term storage of VD versus AD suggests less early impairment of hippocampal function in VD.

Generally, all parameter estimates decline with increasing cognitive severity in both AD and VD, although none of the groups benefited much from temporary retrieval. As dementia severity progresses, the decline along all parameters for both AD and VD suggests two alternative explanations: (1) each disease spreads to involve structures controlling the storage and retrieval functions measured by the previously unaffected parameters; (2) the lesions of each disease do not spread but become locally more severe and disrupt the interaction between the storage and retrieval functions measured by the previously unaffected parameters. The neuropathological changes in AD spread to other structures as well as become more extensive within each structure, whereas in VD it is plausible that damage can remain restricted to subcortical white matter axonal connections. Inspection of lesion volume and distribution in VD using magnetic resonance imaging in conjunction with the GPT model analysis may hold some promise to provide a test to determine whether the general deterioration of all these parameters can result from severe disruption of one part of a network or whether other structures serving storage and retrieval functions as measured by the four parameters in this model must also become affected to show general decline along all parameters of the model.

Roland and Gulyas (1995) have shown that during the storage, retrieval, and recognition of complex geometrical patterns, four structures show increased regional cerebral blood flow using positron emission tomography. These are the posterior inferior temporal lobe, the precuneus, the angular gyrus, and the posterior superior parietal lobule. Damage to any of these structures could affect any or all of the parameters of the model. In AD, these structures do accumulate neuropathological lesions by the middle stages of the disease. We do not know whether these structures are routinely affected in VD, but Sultzer et al. (1995) have demonstrated reduced cortical metabolism with subcortical ischemia and without MRI evidence of cortical involvement.

The neuropsychological interpretation of all of these results must be taken with caution, however, because of factors that may be playing a role in obtaining representative parameter estimates, such as subject and item differences. In the application of cognitive models such as ours to data from normal populations, it has become traditional to group data over subjects and items and to study differences in parameters due to experimental manipulations. In the present
application, subject differences, even within a severity category, may be considerable. Further, as with most neurolological test batteries, item order on each trial is fixed for all subjects rather than counterbalanced, as it would be in a typical experimental study. Both subject and item differences may introduce noise into statistical analyses of the data. These issues for GPT modeling are discussed both analytically and with Monte Carlo studies in a number of publications, for example, Batchelder and Riefer (1986, 1990) and Riefer and Batchelder (1988, 1991). Often it is found that analyses of data sets with GPT models are robust under introducing beta distributions of the models’ parameters, with moderately large variances. This work suggests that the estimates are usually good approximations to the group average parameter values. However, with large numbers of individual differences, confidence intervals and G^2 values based on asymptotic theory such as the ones in Table 4 must be regarded as only suggestive. Thus, model analyses such as ours can be only one part of a converging set of evidence.

**DISCUSSION**

The primary goal of the foregoing analysis was to introduce a relatively simple method for understanding some of the underlying processes in memory recall performance. Using the statistical GPT analysis, the investigators found significant group differences in parameters of immediate free recall performance — for example, between early AD and VD — that have not been reported using traditional statistical techniques. Further, the model analyses allows these group differences to be explained in terms of the changes in underlying cognitive processes, as measured by the parameter estimates.

There have been efforts to disentangle underlying cognitive processes with operationally defined performance scores (e.g., Geffen et al., 1990; Mitrushina & Satz, 1989; Mitrushina et al., 1991). There studies analyzed data from normal elderly subjects using the Rey Auditory-Verbal Learning Test (RAVLT; Rey, 1964), which provides measures of acquisition, storage efficiency, retrieval efficiency, and short- and long-term memory, among other cognitive measures. The task involves 5 immediate study-test free recall trials of 15 items, followed by presentation and free recall of an interference list (Trial 6), a test-only trial of the original items (Trial 7), and a delayed (20-min) recognition test. Acquisition was typically measured by performance on the recognition task or by the difference between the total number of words recalled on Trial 5 minus Trial 1; storage efficiency was operationalized by forgetting rates, which were the total number of items recalled from Trial 5 minus the total from Trial 7; retrieval efficiency involved comparisons of recall and recognition performance; and short- and long-term memory was operationalized by recency and primacy effects, respectively. Each of these studies was successful in detecting differences in one or more of the measures listed above between subject populations (i.e., age groups).

While it may be possible in some cases to operationally define statistics that bear on underlying cognitive processes, this approach is unlikely to work if there are more than a few underlying processes that interact in a nonlinear fashion to create the observable data. For example, storage efficiency in the RAVLT is measured by the forgetting rate from Trial 5 to Trial 7. Although this can be considered a measure of storage capacity, the presentation of an interference list between Trials 5 and 7 introduces an additional factor that can influence recall performance on Trial 7. It is possible that decreased performance on Trial 7 reflects difficulties in retrieval and/or source memory due to proactive interference; subjects may either have trouble inhibiting responses from the interference list or they may not be able to determine which list (original or interference) contains the appropriate items for recall.

Batchelder and Riefer (1986, 1990) and Riefer and Batchelder (1988) have several explicit examples where traditional operational measures of underlying cognitive processes may fail to measure them accurately. GPT models enable one to model nonlinear interactions among underlying cognitive processes. An additional and important advantage of formal modeling over operational definitions is that the formal model enables one to simulate data with various parameter values and thus study the impact of parameter variation on observable statistics. At a minimum we recommend that our approach be used in conjunction with more traditional statistical approaches when it is possible to construct a reasonable GPT model of the task.

It is clear that it will be necessary to have several other lines of converging evidence before the specific conclusions of this study can be accepted with high confidence. In this spirit, perhaps our most important conclusion is methodological, namely, by taking full advantage of the available data, it is possible to delve much deeper into the possible psychological roots of memory deficits in AD and VD. The traditional approach of examining performance score differences as illustrated in the learning curve analysis fails to consider much of the available data.

We close by providing additional compelling evidence that the learning curve analysis is overlooking information in the data. We developed an alternative model of the sequential data in Table 3 that reflected a simple Bernoulli process ("coin flip" on each trial) between recalling and not recalling an item. For example, if \( p_1 \), \( p_2 \), and \( p_3 \) are the recall probabilities on each trial, respectively, then the probability of the sequence 010 would be \( (1-p_1)p_2(1-p_3) \) because the independent Bernoulli trials model assumes that all the information in the data like that of Table 2 is contained in the marginal performance on each trial. The goodness-of-fit test revealed a poor fit, ranging from \( G^2(4, N = 260) = 28.83 \) (for severe VD) to \( G^2(4, N = 1600) = 158.79 \) (for controls), \( p < .05 \). This result points strongly to the fact that much more information is extractable than is obtained using trial means. This fact, of course, is what motivates the examination of the data with a GPT model such as the one reported here.

**ACKNOWLEDGMENTS**

Support for this research came from a National Institutes of Health pilot project grant to William H. Batchelder as part of a larger grant from the National Institute on Aging (P50 AG-05142), with funding dates April 1, 1994 to March 31, 1999. The authors would like to thank David Riefer for helpful insights.

Address correspondence to Dr. William H. Batchelder, Department of Cognitive Sciences, University of California, Irvine, CA 92697. E-mail whbatch@aris.ss.uci.edu
REFERENCES


Received June 3, 1996
Accepted April 29, 1997