January 2013

A Model of Positive Sequential Dependencies in Judgments of Frequency

Jeffrey Scott Annis
University of South Florida, jannis@mail.usf.edu

A Model of Positive Sequential Dependencies in Judgments of Frequency

by

Jeffrey Annis

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Kenneth Malmberg, Ph.D.
Toru Shimizu, Ph.D.
Joseph Vandello, Ph.D.

Date of Approval:
September 29th, 2011

Keywords: recognition memory, model selection, absolute identification, similarity, REM

Table of Contents

List of Tables

List of Figures

Abstract

Positive sequential dependencies occur when the response on the current trial *n* is positively correlated with the response on trial *n*-1. This was recently observed in a Judgment of Frequency (JOF) task (Malmberg and Annis, 2011). A model of positive sequential dependencies was developed in the REM framework (Shiffrin & Steyvers, 1997) by assuming that features that represent the current test item in a retrieval cue carry over from the previous retrieval cue. To assess the model, we sought a set of data that allows us to distinguish between frequency similarity and item similarity. Therefore, we chose to use a JOF task in which we manipulated the item similarity of the stimuli by presenting either landscape photos (high similarity), or photos of everyday objects such as shoes, cars, etc (low similarity). Similarity was modeled by assuming either that the item representations share a proportion of features or by assuming that the exemplars from different stimulus classes vary in the distinctiveness or diagnosticity. The model fits indicated that the best way to model similarity was to assume that items share a proportions of features.

Cognitive testing often assesses the performance on a simple task over the course of many test trials. Many models assume the independence of the individual responses. In fact, many models must make this assumption to be valid instruments for measuring or understanding. That is, the independence assumption is critical for understanding how the tasks are performed, and it is required by many statistical tests, including maximum-likelihood analyses, analyses of variance, etc. (Anderson, 1971). However, there are extensively documented cases in the psychological literature where the independence assumption does not hold.

Perhaps the most well known example of a task in which human cognition violates the independence assumption is absolute identification. In an absolute identification task, the subject classifies stimuli, usually along a single perceptual dimension, and the number of stimulus categories is equal to the number of mutually exclusive responses. For instance, tones of *m* different frequencies must be classified along an *m*-point scale by assigning an integer corresponding to one point on the scale to each stimulus. Absolute identification requires training - and even then it is difficult. The upshot is that errors are made in the classification process, and these errors are non-random; correlations between the responses given in different intervals or sequential dependencies (SDs) are a robust finding (Ward & Lockhead, 1971; Lacouture, 1997; Mori, 1989; Mori & Ward, 1995). Positive SDs occur when the current response is positively correlated with a previous response (or stimulus), which is known as *assimilation*. For example, consider the response to the current stimulus on trial n, S_n, and the response to the previous stimulus value on trial n-1, S_{n-1}, and the difference between successive stimuli, $S_{n-1} - S_n$. As the difference increases, the subjects' estimate of the

current stimulus also increases. Negative SDs or *contrast* in absolute identification is also observed, typically at lags greater than 1 and only when feedback is provided (Ward & Lockhead, 1970).

SDs are also found in episodic memory recognition tasks, including detection and confidence ratings, and judgments of frequency (Malmberg & Annis, 2011). Recognition is the ability to determine what was previously experienced. Usually, subjects in a laboratory experiment study a list of items, and for detection, subjects classify stimuli as having been studied or having not been studied. Assimilation is observed when $S_{n\text{-}1}$ and S_{n} are more likely than chance to be classified as both studied or both unstudied. There are other procedures for testing recognition memory, however. Testing recognition memory via a Judgment of frequency (JOF) is somewhat analogous to absolute identification since it requires a mapping of *m* classes of stimuli to *m* responses, where $m > 2$; items are studied various number of times, and the subject responds with the number of times the word was presented at study. Unlike, absolute identification subjects require no training in order to perform the JOF task, but like absolute identification, the JOF task is challenging, and SDs are observed. Assimilation is observed between the current response and the previous stimulus and the previous response both when feedback is and is not provided (Malmberg & Annis, 2011). Contrast is observed between the previous response and current response at lags of 1 to 3 in the absence of feedback.

Thus, although the decision structure of absolute identification and JOFs are the same, there are obvious and more subtle differences between the tasks. Indeed, the pattern of SDs observed in absolute identification and JOFs are different (Malmberg & Annis, 2011), and thus models designed for absolute identification do not necessarily

generalize to recognition memory tasks. Here, we present the results of an initial investigation of the mechanisms underlying SDs in recognition memory. We will present a process-model of the JOF task that captures the positive SDs in recognition memory. The framework of the model is the retrieving effectively from memory theory (REM, Shiffrin & Steyvers, 1997; Malmberg, Holden & Shiffrin, 2004), and like all memory models, REM assumes that recognition is based on the outcome of an interaction between a retrieval cue and the contents of memory. The retrieval cue is a mental representation of the stimulus, and the more similar the cue is to the contents of the memory the more familiar it seems. For a JOF, we assume that the greater the familiarity of the stimulus the greater the JOF will be that is assigned to it. We will discuss these assumptions in detail below.

A key assumption of the model is that assimilation is the result of a carryover of the information used to probe memory from trial to trial. Hence, to the degree that the features in the retrieval cue on trial *n*-1 carryover to the retrieval cue on trial *n*, the information used to probe memory is correlated from trial to trial during JOF testing. Another key assumption of the model is that the carryover is the result of lapses in attention or vigilance, and therefore the carryover of features does not necessarily occur on each JOF trial. Last, the model assumes that the subject is unaware that carryover occurs, and therefore the subject fails to discount the cross trial correlations in the information gleaned from memory (cf. Huber, Shiffrin, Lyle, & Ruys, 2001).

In prior modeling of JOFs, the similarity of the stimuli has been shown to be critical (Hintzman, Curran, & Oppy, 1992; Malmberg, et al., 2004); since the similarity of the stimulus to the contents of memory is positively correlated with the JOF, JOFs are

greater on average for items more similar to contents of memory, even when they were not studied, than items less similar to the contents of memory. Moreover, noting that the stimuli in a recognition experiment may independently vary in similarity on two dimensions, distinguishing recognition from absolute identification. In absolute identification, the stimuli vary in similarity only along the dimension used to classify them at test. For instance, *m* lines of different length may be classified into *m* categories at test. For the JOF task, not only do the stimuli vary in the number of times that they were studied, but they may also vary on other dimensions, such as perceptual or semantic characteristics. Thus, we conducted an experiment in which repetitions at study were varied and the perceptual/semantic similarity of the stimuli was varied. However, before reporting these results, we will describe the model in greater detail.

A Model of Judgments of Frequency (JOFs)

REM assumes that lexical/semantic traces are represented as vectors of *w* geometrically distributed, features values (Shiffrin & Steyvers, 1997). The environmental base rate of feature values is determined by the geometric distribution parameter *g*. When an item is studied, its lexical/semantic trace is activated, and *t* attempts to store a feature to an episodic trace are made. The probability that a feature will be stored on each attempt is u^*. If the feature is stored, it is copied correctly from the lexical/semantic trace with probability *c,* otherwise the feature value stored is 0. If the feature is not copied correctly, then the stored value is drawn randomly from the geometric distribution:

$$P(V = j) = g(1 - g)^{j-1} \quad 1.$$

where $j \in \{1 \ldots \infty\}$. The predictions of REM do not hinge on the assumption of geometric distribution of feature values. Rather, the geometric distribution is convenient to assume

since the single parameter g defines a given distribution: the mean feature value is $1/g$ and variance is $(1-g)/g^2$. Three geometric probability mass functions are plotted in Figure 1. As g increases, the mean feature value increases and the variability of the feature values of which items are constructed increases. Thus, when the geometric distribution is defined by relatively low values of g, the representations that are created will tend to consist of a wider variety of features values, and the mean feature value will be greater, compared to when the geometric distribution is defined by a relatively low g value. g affects the similarity of the representations.

Repetitions. For the JOF task, items are studied one or more times on a long study list. There are a number of ways to model item repetitions (Criss, Malmberg, & Shiffrin, 2011; Shiffrin & Styevers, 1998). One might assume that a new trace is stored in memory or might assume that previously unstored features are added to the prior trace. However, REM is constrained by the assumption that more often than not item repetitions are encoded by adding previously unstored features to an existing trace. Again there are number of different ways to satisfy this requirement, but the simplest model assumes that each study repetition results in additional storage attempts to the same trace in memory, and since the more complex models do not have an any obvious advantages for the present purpose this is the model we chose to implement (Malmberg & Shiffrin, 2005; Shiffrin & Steyvers, 1997,1998). We will also ignore in the present simulations the effects of the repetition of an item during the course of testing (Malmberg, Criss, Gangwani, & Shiffrin, 2012). Hence, when an item is repeated, unstored features, represented by 0 values, are overwritten in the event of a successful storage attempt and

that existing features are preserved. Therefore, the probability of storing a feature in episodic memory after studying an item *r* times is

$$P(storage) = 1 - (1 - u^*)^{tr} \quad 2.$$

JOFs. During single item recognition, the test item's associated lexical/semantic trace serves as the retrieval cue. The retrieval cue is matched in parallel against episodic traces stored during study. For each episodic trace, *j*, a likelihood ratio, λ_j, is computed:

$$\lambda_j = (1 - c)^{n_{jq}} \prod_{i=1}^{\infty} \left[\frac{c + (1 - c)g(1 - g)^{i-1}}{g(1 - g)^{i-1}} \right]^{n_{ijm}}, \quad 3.$$

where, n_{jq} is the number of non-matching features in *j*, and n_{ijm} is the number of matching features in episodic trace *j*. The first term of Eq. 2 simply represents the contributions of mismatching features to the likelihood ratio, which occur when features are encoded incorrectly during study. The denominator of the second term, represents the chance of obtaining a match of feature value *i*, given that the match was obtained by randomly sampling that feature value from the geometric distribution defined by *g*. Likewise, the numerator of the second term, represents the chance of obtaining a match for feature value *i* given that it was stored correctly during study. Note that the denominator decreases with increases in the feature value, *i*. Hence, matches of relatively great feature values lead to a relatively great likelihood ratio because they are more diagnostic (i.e., less likely to have occurred by chance). This produces greater HRs and lower false-alarm rates for items generated from the geometric distribution with relatively low *g* values. The log odds are obtained from the average likelihood ratio for the *n* traces compared to the retrieval cue,

$$\phi = \ln\left(\frac{1}{n}\sum_{j=1}^{n}\lambda_j\right), \quad 4.$$

where n is the number of episodic traces stored during study, are then compared to a criterion. For binary old-new recognition, when the log odds are greater than 0, an "old" response is given.

Decision Model. The JOF decision mechanism proposed by Malmberg, Holden & Shiffrin (2004) assumes that the log odds on each trial are compared to a set of decision criteria. These criteria are generated according to the equation:

$$\boldsymbol{C}_k = ln(k)r, \quad 5.$$

where $\boldsymbol{C}_k$ is the log odds associated with the JOF k, where $k = 1\ldots n$, and r is a scaling parameter. The log odds are compared to the set of criteria. The JOF corresponds to the value k associated with the greatest criterion exceeded. If the odds do not exceed any criteria, the JOF corresponds to the value k associated with minimum criterion value.

An Important Distinction: Item Similarity versus Frequency Similarity. The dimension on which the JOFs are made is the unidimensional familiarity value, ϕ, obtained from the global-matching retrieval process (Eq. 3). The familiarity of adjacent test items are correlated to the extent that they were presented a similar number of times during the study phase; each condition will produce a distribution over ϕ and the more similar the number of times adjacent tests are studied, the greater the overlap of their respective distributions and the greater the correlation in the adjacent responses at test. We refer to this dimension as *frequency similarity*.

The model also predicts correlations among adjacent responses during testing to the extent that their episodic representations and retrieval cues are comprised of similar

sets of features. Compare this to a typical absolute identification experiment, where the stimuli always vary in similarity only along some perceptual dimension, and this variable will be positively related among adjacent stimuli only to the extent that the stimuli are perceptually similar to each other. In recognition memory testing, however, it is important to distinguish between frequency similarity and the similarity with which items are represented and the similarity of the items, which we refer to *item similarity*. The representation of two items may be very different (e.g., DOG and TRANSITOR), but the information that they elicit from memory would tend to be similar when studied the same number of times. Thus, we would expect there to be stronger correlation between the JOFs given to items presented similar number of times, even though they are represented quite differently, than to an item, say, that was not studied at all.

Modeling Stimulus Similarity Frequency similarity is modeled by varying the number of times an item is presented during study in the manner discussed above. Here, we consider two models of item similarity. The first model assumes that a proportion of feature values were shared among traces in memory. In order to generate similar traces, a vector, $\boldsymbol{P}$, was filled with feature values according to the geometric process outlined above. For each additional vector $\boldsymbol{A}$, $P(\boldsymbol{A}_i = \boldsymbol{P}_i) = s$, where i is the index of the element of the vector. Thus, as the parameter s increases, the proportion of features shared between two representations increases. The second model assumed that the distribution of feature values differed in terms of the g parameter. The g parameter models the environmental base rate of feature values. Note that according to Eq. 1, low base rates correspond to highly distinctive or diagnostic features whereas high base rates generate feature values that are less diagnostic. In terms of computation, as the g parameter increases, the

distribution of feature values becomes positively skewed, as shown in Figure 1. The key distinction between these two models is that the latter model assumes that less similar stimuli not only overlap less in the representations, but less similar representations are also comprised of more diagnostic or uncommon features and are more distinctive representations (Malmberg, Steyvers, Stephens, & Shiffrin, 2002; Shiffrin & Steyvers, 1997). Note that these models need not be mutually exclusive and we will therefore also consider the more complex model in which both *s* and *g* are used to create systematic variability in item similarity.

Assimilation. We refer to this model of assimilation as the *carry over model*. It shares a key assumption with several models of absolute identification; the information on which a decision is made on trial *n*-1 is not independent of the information in which the decision on trial *n* is made (Brown, Marley, Donkin, & Heathcote, 2008; Petrov & Anderson, 2005; Stewart, Brown, & Chater, 2005; Triesman & Williams, 1984). More specifically, we speculate that the degree to which information carries over from trial to trial during the course of testing may fluctuate due to the ability of the subject to maintain an optimal level of vigilance when performing the task. On each trial, there is a probability, 1-*a*, that a carryover process occurs in which each feature from the retrieval cue on trial *n* carries over to the retrieval cue on trial *n*+1 with probability *b.* Therefore, on each trial with probability *a*, no carryover process occurs. For example, if *a* and *b* were both equal to 1, no carryover would occur because the system would essentially "refresh" itself on each trial. Therefore, as *a* increases, the number of trials in which carry over occurs decreases. When the refresh parameter is equal to 1, the model reverts to the Shiffrin and Steyvers, (1997) model.

Another key assumption made explicitly here but tacitly apart of other models is that the subject fails to discount the information carrying over from trial-to-trial (cf. Huber, Shiffrin, Lyle, & Ruys, 2001). The failure to discount the carry over produces assimilation. For example, imagine that a subject carries over all the features from trial 1 to trial 2. They then globally match the features in the retrieval cue to the contents in memory (see Equation 1) and generate an odds value (see Equation 2). In this case, the odds value on trial 2 is equal to the previous odds value on trial 1. Thus, the subject makes the same response on trial 2 as he or she did on trial 1. Let us say, on trial 3, however, the subject refreshes their retrieval cue and does not carry over any features from trial 3. In this case, the subject is free from any influence of previous trials and makes a response that is independent of all other responses. Therefore, if the subject refreshes their retrieval cue on every trial, all responses would be independent. This is the assumption made in the original Shiffrin and Steyvers (1997) model and would only occur in the current model when the refresh parameter takes on the value of 1.

Experiment

The carry over model creates positive sequential dependencies in recognition memory testing because the information on which a decision is made on from trial to trial is correlated. However, it is unclear whether the carry over model can provide a qualitatively accurate account for assimilation among JOF responses. To assess the carry over model, we sought a reasonably complex and challenging set of data that allows us to distinguish between frequency similarity and item similarity. For this reason, we presented subjects with items that were either high or low in item similarity with photos of landscapes (e.g. mountains, sunsets, fields etc.) corresponding to high similarity items,

and photos of everyday random objects (e.g. shoes, chairs, cars etc.) corresponding to low similarity items. Figure 2 shows a sample of the items presented. To manipulated frequency similarity, the items in both conditions were presented from 1 to 6 times during study, and JOFs were collected to test recognition memory.

Method

Subjects. One-hundred-and-ten undergraduate students at the University of South Florida participated in exchange for course credit.

Design and Materials. Repetitions were manipulated within subjects and within lists, and similarity was manipulated between subjects. Similarity of the stimuli were manipulated by presenting either landscape photos (high similarity), or photos of everyday objects such as shoes, cars, etc (low similarity). The 240 color object images consisted of everyday inanimate objects such as shoes, chairs, motor vehicles, clocks, food, kitchenware, candles etc., while the 240 color landscape images consisted of sunsets over beaches, mountains, parries, etc. Four lists of 60 images each were studied. The images were drawn randomly and anew for each subject from the 240 images described above. Within each list, 10 images were presented for 1.0 s, either once, twice, three, four, five, or six times with at least 1 intervening image between each repetition. Each test list consisted of the 60 images presented at study. 55 subjects completed the object condition, and 55 subjects completed the landscape condition.

Procedure. Subjects studied four lists of images and performed a math task after each. The math task consisted of mentally adding digits for 30-seconds. Upon completion of the math task, each image from the study list was presented one at a time, and the

subject's task was to indicate how many times the word was studied by typing the appropriate number into the computer using the numerical keys 1-6.

Results

Proportion of Correct Responses. A 2 (stimulus type: objects vs. landscapes) x 6 (number of presentations) omnibus ANOVA was conducted with stimulus type as a between subjects factor and the number of presentations as a within subjects factor. The left panel of Figure 3 shows the proportion of correct responses was greater in landscape condition than in the object condition, $F(1,108) = 122.57$, $MSE = .03$, $p < .0005$, $\eta_p^2 = .53$. There was a main effect of the number of presentations on accuracy, $F(5,540) = 104.10$, $MSE = .161$, $p < .0005$, $\eta_p^2 = .49$, and the stimulus type interacted with the number of presentations, $F(5,540) = 8.19$, $MSE = .02$, $p < .0005$, $\eta_p^2 = .07$. To investigate the interaction, a trend analysis was conducted. The linear trend for the object condition was significant, $F(1, 54) = 56.74$, $MSE = .05$, $p < .0005$, $\eta_p^2 = .51$, as well as in the landscape condition, $F(1,54) = 81.37$, $MSE = .03$, $p < .0005$, $\eta_p^2 = .60$. The quadratic trends for the object condition, $F(1,54) = 138.47$, $MSE = .05$, $p < .0005$, $\eta_p^2 = .72$, and landscape condition were also significant, $F(1,54) = 27.12$, $MSE = .03$, $p < .0005$, $\eta_p^2 = .33$. As the number of presentations increases, the proportion of correct responses decreases in a nonlinear fashion until the number of presentations equals 6, where the proportion of correct responses again increases. However, the characteristic trough for middle range of frequencies was shallower in the landscape condition than in the object condition.

Mean JOFs. The calibration curve in the middle panel of Figure 3 shows a main effect of stimulus type on the mean JOF, $F(1,108) = 19.33$, $MSE = 1.95$, $p < .001$, $\eta_p^2 =$

.15. There was a main effect of the number presentations on the mean JOF, $F(5,540) = 527.71$, $MSE = .14$, $p < .0005$, $\eta_p^2 = .83$, and a stimulus type x number of presentations interaction, $F(5,540) = 98.67$, $MSE = .14$, $p < .0005$, $\eta_p^2 = .48$. Thus, the subjects in the landscape condition tended to, on average, overestimate low stimuli and underestimate high stimuli more often than those subjects in the object condition. According to a linear trend analysis, the mean JOF in the object condition increased with increases in the number of presentations, $F(1,54) = 525.08$, $MSE = .69$, $p < .0005$, $\eta_p^2 = .91$. The quadratic trend was also significant, $F(1,54) = 55.08$, $MSE = .69$, $p < .0005$, $\eta_p^2 = .51$. Mean JOFs increased linearly with increases in the number of presentations in the landscape condition as well, $F(1,54) = 234.77$, $MSE = .25$, $p < .001$, $\eta_p^2 = .81$; however, the quadratic trend was not significant, $F < 1$. The calibration curves suggest that subjects had more difficulty discriminating the number of times that landscapes were presented than the number of times the objects were presented.

JOF Accuracy. To measure JOF accuracy, $d'_{i, i+1}$was calculated for each stimulus i and i+1 (Luce, Nosofsky, Green, & Smith, 1977), which measure the ability of the subject to discriminate items that are adjacent to each other on the frequency dimension in manner independent of range restrictions or response bias. For instance, $d'_{2,3}$ is measure of the ability of the subject to discriminate between items presented two times and items presented three times. The right panel of Figure 3 plots $d'_{i, i+1}$ as a function of the number of presentations, I, and the stimulus condition. A 2 (stimulus type: objects vs. landscapes) x 5 (number of presentations i) omnibus ANOVA revealed a main effect of the stimulus type, $F(1,108) = 101.30$, $MSE = .30$, $p < .0005$, $\eta_p^2 = .48$, and number of presentations i , $F(4,432) = 21.08$, $MSE = .14$, $p < .0005$, $\eta_p^2 = .16$, on $d'_{i, i+1}$. JOFs were

more accurate for objects than landscapes. Moreover, while $d'_{i,\, i+1}$ steeply declined with increases in the number of presentations in the object condition, a flat curve was observed in the landscape condition, $F(4,432) = 18.29$, $MSE = .14$, $p < .0005$, $\eta_p^2 = .15$. Comparing the right panel to the left panel of Figure 3, one therefore concludes that much of the bow observed in the proportion of correct responses in the landscape condition is due to range restrictions affecting biases in decision making, whereas the bow observed in the object condition is associated with changes in the ability to discriminate between the number of times that the objects were presented.

Assimilation. The left panel of Figure 4 plots the mean error on trial *n* as a function of the current stimulus and prior response. A 2 (stimulus type) x 6 (current stimulus) x 6 (previous response) omnibus ANOVA was conducted. There was no main effect of the stimulus type on the mean error on trial *n*, $F < 1$. There was a main effect the current stimulus, $F(5,265) = 322.19$, $MSE = 1.01$, $p < .0005$, $\eta_p^2 = .42$; as the number of times a stimulus was studied decreased, the overestimate of the JOF increased. There was also a main effect of previous response on the mean error on trial *n*, $F(5,265) = 36.88$, $MSE = .43$, $p < .0005$, $\eta_p^2 = .41$; as the JOF given on the prior trial increased, the JOF on the current trial tended to increase. That is, positive SDs (i.e. assimilation) were observed toward the previous response. There was a significant interaction between the current stimulus value and the stimulus type, $F(5,265) = 38.66$, $MSE = 1.01$, $p < .0005$, $\eta_p^2 = .42$. This reflects the overall increase in accuracy in the object condition versus the landscape condition. There were no other significant interactions.

The right panel of Figure 5 plots the mean error on trial *n* as a function of the current and prior stimulus. A 2 (stimulus type) x 6 (current stimulus) x 6 (previous

stimulus) omnibus ANOVA was conducted. There was a main effect of stimulus type on the mean error on trial *n*, $F(1,89) = 14.85$, $MSE = 12.84$, $p < .0005$, $\eta_p^2 = .14$, such that the overall mean error in the object condition, ($M = -.24$), was higher than the landscape condition, ($M = -.72$). There was a main effect of the number of presentations, $F(5, 445) = 728.18$, $MSE = .89$, $p < .0005$, $\eta_p^2 = .89$, and a main effect of the prior stimulus on the mean error on trial *n*, $F(5,445) = 6.02$, $MSE = .42$, $p < .0005$, $\eta_p^2 = .06$. There were no significant interactions.

The previous response and the previous stimulus are confounded (Jones, Love and Maddox 2006). One simple way to decorrelate the previous stimulus with the previous response is to hold the previous stimulus constant and only let the previous response vary. In order to make sure there would be a sufficient amount of data to conduct the analysis we binned the responses such that responses 1 and 2, 3 and 4, and 5 and 6 would each constitute a bin. We conducted a 2 (stimulus type) x 3 (current stimulus) x 3 (previous stimulus) x 3 (previous response) omnibus ANOVA in order to account for these effects. We again found no main effect of the stimulus type, $F(1,25) = 3.99$, $MSE = 3.99$, $p = .057$, $\eta_p^2 = .14$, a main effect of the current stimulus, $F(2,50) = 257.47$, $MSE = 5.50$, $p < .0005$, $\eta_p^2 = .91$, a main effect of the previous response, $F(2,50) = 34.22$, $MSE = .79$, $p < .0005$, and no main effect of the previous stimulus, $F < 1$. There was a previous stimulus by current stimulus interaction, $F(4,100) = 3.13$, $MSE = 1.61$, $\eta_p^2 = .11$, such that as the previous and current stimulus value increased, the error on the current trial became more negative. Thus, when the previous stimulus and response are decorrelated, negative SDs toward the previous stimulus were observed.

Modeling Results

Fitting Methods. In order to test whether differences existed in parameter values across conditions, a bootstrap method was used. In each bootstrap procedure, the model was fit to the data and parameter estimates were obtained. Using these parameter estimates, data was then generated from the model itself. The model was then fit to this data and new parameter estimates were obtained. The simulations were run on USF's Beowulf Cluster. Each computer on the cluster had a either a Xeon X-series processor or an Opteron 2000-series processor and 16 to 24 gigabytes of memory.

In each bootstrap, we independently varied either *g*, *s* or *a*, between the stimulus conditions. All other parameters were fixed (see Table 1). In this sense, the fits of that we report are not necessarily the best fits possible. However, since the majority of the values used for the fixed parameters were determined a priori from the results of great deal of other simulation previously reported (e.g., Malmberg, et al., 2004; Shiffrin & Steyvers, 1997) the results can be viewed as being relatively parsimonious. We should also note that the parameters for the decision model were fixed, as we were interested in only assessing the ability of the carry over model to account for assimilation in JOF testing and there was no obvious a priori reason to believe that the decision parameters would vary between stimulus conditions. We have more to discuss on this matter in General Discussion.

g, *s*, and *a* were allowed to vary between 0 and 1, except for the *g* parameter, which was only allowed to vary between .1 and 1. This was done to avoid division by zero in equation 1. The downhill simplex method (Nelder & Mead, 1965) was used to estimate the global maximum of the likelihood function. Because the data were averaged,

we assumed that the data generating process was Gaussian in nature and all data points were independent. Therefore, the normal likelihood function was appropriate.

$$L(x_i|u,\sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\frac{(x_i-u)^2}{\sigma^2}\right), \quad 6.$$

where x_i is the *i*th data point, u is the mean generated from the model, and σ^2 is the variance. The maximum likelihood estimate of σ^2 is

$$\sigma^2{}_{ML} = \frac{\sum_i (x_i-u)^2}{n}, \quad 7.$$

where n is the sample size. When substituted for σ^2 the log-likelihood function becomes:

$$lnL(x_i|u,\sigma^2) = \left[\frac{1}{\sum_i (x_i-u)^2}\right]^{\frac{n}{2}} \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} exp\left(-\frac{n}{2}\right). \quad 8.$$

Thus, the log-likelihood function was a function of the sample size, and the sum of the squared error (Glover, & Dixon, 2004). Because REM is a stochastic model it was unlikely that simplex procedure would converge. Thus, to ensure a stopping point for the simplex procedure, a maximum of 100 objective function evaluations were allowed. Each evaluation consisted of 50 simulated subjects which completed 5 lists. After the parameter estimates, $\hat{\theta}$, were generated from the simplex method, 50 bootstrap samples were generated using $\hat{\theta}$. The model was then fit to each of the bootstrap samples in order to generate a distribution of parameter estimates $\hat{\theta}^b$.

In order to evaluate the goodness-of-fit, the Likelihood Ratio Test (LRT) was used according to equation 8.

$$G^2 = -2\ \ln L_{specific} - (-2\ \ln L_{saturated}). \quad 9.$$

The G^2statistic is distributed with *K* degrees of freedom where *K* refers to the number of extra parameters in the saturated model. $L_{specific}$ is the log-likelihood of the model in question, and $L_{saturated}$ is the log-likelihood of the saturated model.

Noting the drawbacks of Null Hypothesis Significance Testing (Wagenmakers, 2007), Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion were calculated (BIC; Schwarz, 1978).

$$AIC = -2\,lnL + 2K\,. \tag{10.}$$

lnL is the log-likelihood and *K* is the number of free parameters.

$$BIC = -2\,lnL + K\,lnN. \tag{11.}$$

The first term represents the negation of the log-likelihood, and the second term is a penalty term, where *K* is the number of free parameters and *N* is the number of observations. The model with the lowest BIC is the model with the highest posterior probability. The BIC does not provide a measure of anything by itself, rather it needs to be compared to another BIC. These differences can be better interpreted by calculating the ratio of the posterior probabilities to obtain the Bayes factor (Kass & Raftery, 1995):

$$B = \frac{p(M_1|D)}{p(M_2|D)} \tag{12.}$$

The Bayes factor is a function of the BIC because $-2ln(p(M|D)) \approx BIC$. Therefore,

$$-2\,lnB \approx [BIC_1 - BIC_2]\,. \tag{13.}$$

Dividing both sides of the equation by -2 and cancelling the log function, equation 12 can be rewritten as

$$B \approx exp\left(\frac{BIC_2 - BIC_1}{2}\right). \tag{14.}$$

Thus, the Bayes factor was approximated by equation 13. The Bayes factor was then used to calculate the posterior probability of each model given the available pool of models (Equation 15).

$$w = \frac{B_M}{\sum_i B_i}. \quad 15.$$

Bootstrap Results. The results of the bootstrap procedure are shown in Tables 2 through 6. In the first simulation, either *g*, or *a*, or *s* was varied. Table 2 shows that the *g* parameter significantly increased from the object (M = .362, SD = .009) to the landscape condition (M = .409, SD = .008), $t(49) = 98.52$, $p < .05$. Thus, the model predictions were in accordance with intuition; the distinctiveness of the features decreased from the object to the landscape condition. Likewise, the *s* parameter increased from the object (M = .431, SD = .034) to the landscape condition (M = .495, SD = .018), $t(49) = 11.94$, $p < .05$. The *a* parameter decreased from the object (M = .777, SD = .009) to the landscape condition (M = .567, SD = .008), $t(49) = -123.14$, $p < .05$. Thus, the model suggests an increase in the number of trials that the carryover process occurred on, from the object to landscape condition.

In order to account for the possibility that a combination of varying parameter values would provide a better fit to the data, we considered the models in which the *g* and *a* parameters were allowed to simultaneously vary (Table 3). A distribution of parameter estimates from a bootstrap procedure was again generated. The *g* parameter was shown to significantly increase from the object (M = .421, SD = .004) to the landscape condition (M = .444, SD = .002), $t(49) = 37.62$, $p < .05$. The *a* parameter was shown to simultaneously decrease from the object (M = .889, SD = .012) to the landscape condition (M = .639, SD = .004), $t(49) = -113.46$, $p < .05$. Thus, similar model predictions were

observed in that the proportion of trials in which carryover occurred increased from the object to the landscape condition.

We also varied both the *s* and *a* parameters simultaneously and obtained the parameter estimates from the bootstrap procedure described above (Table 4). The *s* parameter significantly increased from the object (M = .285, SD = .070) to the landscape condition (M = .584, SD = .008), $t(49) = 29.90$, $p < .05$ and the *a* parameter significantly decreased from the object (M = .861, SD = .008) to the landscape condition, (M = .491, SD = .012), $t(49) = -195.26$, $p < .05$. Thus, the model predictions are in accordance with the results from the bootstrap procedure in which each parameter was varied independently. That is, the similarity of feature values increased from the object to the landscape condition, and the proportion of trials in which carry over occurred increased from the object to the landscape condition.

Table 5 shows the results of the bootstrap in which *g* and *s* were varied simultaneously. The bootstrap revealed that the *g* parameter increased from the object (M = .396, SD = .003) to the landscape condition (M = .442, SD = .005), $t(49) = 61.41$, $p < .05$, while the *s* parameter increased from the object (M = .546, SD = .014), to the landscape condition (M = .593, SD = .014), $t(49) = 16.93$, $p < .05$. Thus, while the distinctiveness of features decreased from the object to the landscape condition, the overall similarity of traces in memory increased.

Finally, we simultaneously varied all parameters of interest: *s*, *g,* and *a*. The results are shown in Table 6. This model showed a similar pattern in that the *a* parameter decreased from the object (M = .993, SD = .005) to the landscape condition (M = .609, .026), $t(49) = 98.70$, $p < .05$, and the *s* parameter increased from the object (M = .362, SD

= .051) to the landscape condition (M = .622, SD = .004), $t(49) = 34.84$, $p < .05$. However, the g parameter increased from the landscape (M = .433, SD = .005) to the object condition (M = .453, SD = .002), $t(49) = 23.91$, $p < .05$.

The counterintuitive result from the final simulation is worth discussing at greater length. Note, that the simulation suggests that the landscapes were MORE distinctive than the objects when measured by g but less distinctive when measured by s. This result raises red flags. Moreover, when the simpler models were simulated, the results consistently suggested that the landscapes were LESS similar than the objects when measured by both g and s, which is what intuition tells us should be the case. Hence, our suspicion was that the complexity of the a, g, s model is unwarranted, and the counterintuitive parameter estimates that it produced were the result of over fitting (Pitt & Myung, 2002). If so, we suspected that despite the awkward parameterization of the a, g, s model, it would provide a superior quantitative fit of the data, which we explored next.

Model Fits and Selection

Overall quality of fits. The parameter estimates from the different simulations are listed in Tables 2 through 6. Quantitative fits were calculated as described above, and various statistics obtained from the "best" fits of each model are shown in Table 7. The best fits of the data were obtained from the carryover models in which the amount of carryover was free to vary between conditions. According to these models, attention fluctuates at test, and when vigilance is reduced there is a carryover of a proportion of the features from the retrieval cue used to probe memory on trial n-1 to the retrieval cue used to probe memory on trial n. The variability in the attention (a) parameter accounts for the

differences in the magnitude of the SDs observed between the stimulus conditions, and all the bootstrap simulations indicate that that less carryover occurred in the object condition than in the landscape condition. Thus, the most accurate models of recognition that we have considered take into account the sequential dependencies observed in JOF testing, and by taking into account these SDs, the models are better able to account for differences between the stimulus conditions. The result is not as trivial as may seem since the stimulus conditions were constructed in manner that a priori would have been thought to differ along the stimulus similarity dimensions (*g* and/or *s* dimensions), but the differences in the amount of carryover between stimulus conditions characterized all three of the best fitting models.

Of the three carryover models, the *a*, *g*, *s* model and *a*, *s* model provided the best quantitative accounts, suggesting that there was differences in the nature of the stimuli themselves in addition to the differences in the amount of carryover between stimulus conditions. Since the two best fitting models differ in complexity, we inspected the Bayesian Information Criterion (BIC; also Table 7). The model in which *a*, *g*, and *s* were varied simultaneously obtained a slightly lower BIC value than the *a*, *s* model, indicating that even when the complexity of the models is taken into account, the *a*, *g*, *s* model is preferred. In addition, the posterior probability of 1.0, denoted by *w* in Table 7, indicates that the *a*, *g*, *s* model by far and away is the most likely model to have generated the data.

All these statistics are informative, but it is critically important that the parameter estimates obtained from the simulations also make sense. Note, the parameters estimated obtained from the *a*, *g*, *s* model suggests that the landscapes are less distinctive than the objects. This is difficult to reconcile with a casual inspection of the stimuli (see Figure 2).

This is a rather peculiar state of affairs, compounded by the fact that the posterior probability clearly favors the *a, g, s* model over the *a, s* model. However, the advantage for the *a, g, s* model may be more apparent than real. For example, both the *a, g, s* model and the *a, s* model yield similar qualitative fits (see Figures 6 through 9). Moreover, the LRT revealed the *a, g, s* model and *a, s* model were unable to provide an adequate overall fit to the data, $G^2(163) = 762.71$, $p < .05$, and $G^2(164) = 793.44$, $p < .05$, respectively. On the other hand, it is unclear how much to make out of the LRT analysis, since when the number of degrees of freedom is high (that is, when there are many more data points than free parameters), even small departures of the fit from the data become accentuated in the LRT. Although there is no formal means for combining the outcome of a series of simulations with intuition (cf. Shiffrin & Nobel, 1995), the counterintuitive nature of the *a, g, s* model and its additional complexity, suggests that the slight quantitative advantage for the *a, g, s* model over the *a, s* model is due to overfitting (Myung & Pitt, 2000). We therefore, conducted a series of more fine-grained analyses of the model fits to different subsets of the data.

Landscape versus Object Stimulus Conditions. Note these analyses are on the fits obtained from simulations of either all of the object data or all of the landscape data. Hence, although we are assessing the capability of the model to account for various subsets of the data within each stimulus condition, the fits on which the following statistics are derived are constrained by either all of the object data or all of the landscape data. In this sense, the following assessments are overly conservative in judging the models' capacities to account for the fine-grained details of the data. On the other hand,

the data themselves are not independent, as the SDs clearly indicate, and therefore it seemed reasonable to maintain the additional constraints.

We first looked at how the models fit the overall landscape data. Both models make similar qualitative predictions (see Figures 6 through 9). Although both the *a, g, s* model and *a, s* model were unable to quantitatively fit the landscape data, $G^2(80) = 284.58$, $p < .05$, and $G^2(81) = 259.19$, $p < .05$, respectively, we were nevertheless interested in whether one model was able to provide a "better account" of the landscape data than the other. Accordingly, we computed the BIC for each of the models and found the *a, s* model had a lower BIC (7827.86) than the more complicated *a, g, s* model (BIC = 7858.43), and the *a, s* model was more likely to have generated the landscape data ($w = 1.0$). We next analyzed the fits of each model to the object data. Again, both the *a, g, s* model and *a, s* model were unable to fit the object data, $G^2(80) = 478.13$, $p < .05$, and $G^2(81) = 534.25$, $p < .05$, respectively. The BIC of the *a, g, s* model was 8682.46 while the BIC of *a, s* model was 8733.25, and the *a, g, s* model was much more likely to have generated the object data ($w = 1.0$). Thus, while the *a, s* model had a higher posterior probability for landscapes, the *a, g, s* model had a larger posterior probability for objects. This suggests that the superior - but rather peculiar - fit of the *a, g, s* model is due to overfitting the object data.

Percent Correct Responses. To assess the strengths and weaknesses of the different models, we next considered how each model faired within each subset of data in the landscape and object conditions. Both models did a reasonable job at fitting the characteristic bow effect in the landscape and object conditions, and both models predicted increased proportion correct in the object condition (see Figure 6). There are,

however, several aspects of the fit of the proportion of correct responses as a function for frequency to note. First, neither the *a, s*, or *a, g, s* model fit the object percent correct data, $G^2(4) = 4.35$, $p < .05$, and $G^2(3) = 2.86$, $p < .05$, respectively. Comparing the two fits, the *a, s* model has a more difficult time predicting the effect of the stimulus when the repetitions of the items was low in the object condition, whereas the *a, g, s* model handles these data better. Although the *a, g, s* model appears to fit the object bow curve data slightly better, this advantage comes at the price of increased complexity. We therefore calculated the BIC for each model given the object bow curve data. The BIC obtained from the *a, g, s* model was 629.93, while the *a, s* model had a BIC of 627.41. Moreover, the *a, s* model had a greater posterior probability (.779) for the object bow curve data compared to the *a, g, s* model (.221). For the landscape bow curve data, the *a, s* model fits the proportion of correct responses as a function of frequency better for low stimulus values, while the *a, g, s* model does a better job for larger stimulus values. Although the qualitative advantage of either model is not readily apparent via the "eyeball test", the *a, s* model was able to fit the landscape bow curve data, $G^2(4) = .256$, $p > .05$, while the *a, g, s* model was not, $G^2(3) = .60$, $p < .05$. In addition, the posterior probability of the *a, s* model, $w = .898$, was much higher than the *a, g, s* model, $w = .102$. Thus, the *a, s* model was much more likely to have generated the bow curve data than the *a, g, s* model.

Thus, the *a, s* model is the winning model for both the object and landscape bow curve data according to our Bayesian analysis. Whatever advantages the *a, g, s* model has in its goodness of fits appear to be due to overfitting. To make some conclusions about the departure of the models from the data, it is worth noting a couple of things. First, the both models are under-predicting the effect of the stimulus when at relatively low levels

of frequency. Second, we have not allowed the models to vary between stimulus conditions in terms of their decision bias. Third, the percent correct measure confounds the ability of to discriminate along the frequency dimension and bias due to range restrictions (Luce et al., 1982). Therefore, in order to understand why the models under consideration that do not vary in bias between stimulus conditions, it would be useful to assess them against a measure that provides a more accurate determination of the ability to discriminate between levels of frequency.

JOF Accuracy. Why is the *a, g, s* model overfitting the data? Note that the measure proportion correct confounds accuracy with bias (Luce et al., 1982), and that bias was not allowed to vary between the stimulus conditions. Therefore, this lack of flexibility may be the source of some of the problems the models have in fitting proportion of correct responses as a function of frequency. Therefore, we fit the models to the data plotted in Figure 7 in which accuracy, $d'_{i,\,i+1}$, is a function of stimulus type and the number of presentations. Here, we see both models better predict the stimulus effect on accuracy over the ranges of number of presentations.

That having been said, one might ask, "How well are the models fitting the accuracy data?" While neither of the models were able to quantitatively fit the object $d'_{i,\,i+1}$ curve (both p's <.05, see Figure 7), the *a, g, s* model had a larger posterior probability, (w = .970), than the *a, s* model, (w = .030). The *a, s* model, on the other hand, was more likely to have generated the landscape accuracy function, (w = .965), than the *a, g, s* model, (w = .035). More importantly, note that that accuracy is quite low in the landscape condition, and it is poorer than the accuracy in the object condition. Yet, the *a, g, s* model parameter estimate indicates that the landscapes are *more* distinctive than the objects. The

a, g, s model is compensating for the *a, s* model's adequate fits of the low frequency data in the object condition by making all of the objects relatively less distinctive. Less distinctive stimuli lead to lower levels of familiarity, and hence items are more likely to be judged to have been studied a relatively few numbers of times. Therefore, it again appears that varying the *g* parameter in the *a, g, s* model is serving as a proxy for a change in decision bias between stimulus conditions. That is, variability in *g* is allowing the underlying distribution of familiarity values to shift between stimulus conditions in a manner that mimics a shift in response bias. This allows for a better account of the stimulus effect on the proportion correct (in Figure 6), albeit in an inappropriate manner, and therefore the apparent qualitative advantage for the *a, g, s* model is due to overfitting.

This analysis suggests that the major source of the departure of the model from the accuracy data is in the model of the repetitions or frequency similarity, and not the model of sequential dependencies. Both the *a, s* and *a, g, s* models are underpredicting the ability of the subject to discriminate between items presented 4, 5, and 6 times because of the overly simplified model of encoding that we assumed. Accordingly, every time an item is repeated, features are accumulated in the trace stored when the item was first presented. Moreover, inaccurately encoded features are never corrected on later presentations. Relaxing either or both of these assumptions would allow the model to predict greater differences in the mean familiarity values of the items presented relatively frequently. We have addressed these possibilities in a formal manner elsewhere (Criss et al., 2011; Malmberg, et al., 2004). However, the price would be a more complex model of encoding, and since our primary concern is assessing the SDs in recognition testing, we chose not to introduce unnecessary complexities for the time being.

Sequential Dependencies. The carryover of features from one retrieval cue to the next on some trials is the source of positive SDs in the model. Since a robust pattern of SDs was found in both conditions, it seems almost trivially important that the carryover is important. What is critical to note is that in the present model fits, the predictions came from models in which *a* was free to vary between stimulus conditions. In those models in which *a* did not vary, it was set to a modest value (see Table 1) determined to provide a reasonable account of all of the data during a preliminary simulation. Thus, these fits are being used to determine whether variability in carryover between stimulus conditions provides a better account of the data than a model in which a constant amount of variability is assumed for both stimulus conditions.

The top panels of the Figure 8 show both models yield similar qualitative predictions for assimilation towards the previous responses in the landscape condition. The *a, s* model does slightly worse in the object condition, as it overestimates the error for low repetition stimuli. However, for high repetition stimuli, both models make a similar underestimation. Quantitatively, the LRT revealed the *a, s* model and the *a, g, s* model both deviated significantly from the response assimilation data in the landscape condition, $G^2(34) = 162.08$, $p < .05$, and $G^2(33) = 112.13$, $p < .05$, respectively. The BIC and resultant posterior probability for the *a, g, s* model for the landscape response assimilation data was lower ($BIC = 1926.14$, $w = 1.0$) than for the *a, s* model ($BIC = 1918.80$, $w < .0005$). Hence, the more complex *a, g, s* model provides a better account of the positive sequential dependencies observed in the landscape condition, and it is more likely to have generated the data than the *a, s* model. However, note that in order to

generate this superior fit, the *a, g, s* model must assume that landscapes are MORE distinctive than objects.

The bottom panels of Figure 8 plot the fits of the models to the response assimilation data for objects, and the *a, s* model appears to fit the pattern of assimilation slightly better than the *a, g, s* model. However, both the *a, s* and *a, g, s* models failed to quantitatively fit the object response assimilation data, where, $G^2(34) = 221.08$, $p < .05$, and $G^2(33) = 232.16$, $p < .05$, respectively. The *a, g, s* model is having difficulty predicting response assimilation in the object condition because of the large proportion of trials in which the retrieval cue is "refreshed" and no carry over occurs. This is due to the high value of *a* parameter estimate. Quantitatively, this is reflected in the lower posterior probability of the *a, g, s* model, ($w = .001$), while the posterior probability of the *a, s* model was .999. Hence, the simpler *a, s* model provides a better account of the positive sequential dependencies observed in the object condition, and it is more likely to have generated the data than the *a, g, s* model.

The top panels of Figure 9 show the fits of the model to the error plotted as a function of the previous stimulus value for the landscape condition. The quantitative model fits significantly deviated from the saturated model for the *a, g, s* model $G^2(33) = 163.03$, $p < .05$, and *a, s* model $G^2(34) = 162.08$. While both models show a reasonable qualitative fit for this subset of data, the *a, s* model qualitatively captures the overall error magnitudes for each stimulus value better than the *a, g, s* model. This fact is reflected in the BIC values. For this particular data subset, the *a, s* model had a BIC of 4698.57 while the *a, g, s* model had a BIC of 4774.80 which yielded a much higher posterior probability for the *a, s* model ($w = 1.0$).

The bottom panels of Figure 9 show the model fits to the mean error on trial *n* as a function of the previous stimulus value for the object condition. Both models qualitatively predict the absence of assimilation towards the previous stimulus value. However, both models were unable to quantitatively predict this subset of the data, ($p <$ .05 for both models). The *a, s* model overestimates the error magnitude of low and high repetition stimuli more so than the *a, g, s* model. In addition, the *a, g, s* model was clearly the quantitative winner; it had a lower BIC than the *a, s* model (4201.70 vs. 4252.72) and a higher posterior probability ($w = 1.0$).

In summary, the *a, s* model was more likely to have generated the landscape data, but the *a, g, s* model was more likely to have generated the object data. Again, it appears that *a, g, s* model is fitting the data better due to a variability in *g* serving as proxy for variability in decision bias between stimulus conditions.

Similarity versus Distinctiveness. We have considered models in which variability in the overlap of features between stimulus conditions is modeled by changes in *s*, the similarity parameter, and *g*, the distinctiveness parameter in REM. The model allowing both *s* and *g* to vary between stimulus conditions provides superior fits of the data. According to the Bayesian analyses that we conducted, the *a, g, s* model almost certainly generated the data from our experiment given the models that we considered. However, the variability in *g* comes at the cost of counterintuitive parameter estimates that suggest that *g* is standing in place for changes in response bias. Here, we conducted a similar Bayesian analysis, but we did not include the *a, g, s* model in the competition. The question was, did one model provide a better account of the overlap in features between stimulus conditions? The results are presented in Table 8. The *a, s* model

provided the best quantitative account of the data as far as being more likely to have generated the data than the *a*, *g* model.

General Discussion

We described a model of positive sequential dependencies that assumes information from the retrieval cue may be carried over and combined to form the retrieval cue used to probe memory on next recognition trial. The model is an extension of a JOF model used to account for the interactions of normative word frequency, item similarity, and repetitions observed in recognition testing (Malmberg et al., 2004), which assumed that word-frequency was correlated with the distinctiveness of the features used to represent words (Malmberg, Steyvers, Stephens & Shiffrin, 2003; Shiffrin & Steyvers, 1997), and that item similarity is varied by manipulating the proportion shared features among items constructed from a given base rate distribution of feature values.

In the present experiment, we manipulated the nature of the stimuli used to test recognition memory in order to assess the ability of the model to account for positive SDs in JOFs. The stimulus manipulation is provocative within the framework of the carryover model. One question was whether the effect of the inter-item similarity manipulation could be captured simply by variability in the carryover of features used to probe memory at test, the same mechanism used to produce positive sequential dependencies. That is, could we capture the differences in the stimulus conditions by simply accounting for the differences in the SDs? On the other hand, it was unclear whether inter-item similarity should be modeled by varying the overlap of the features used to represent the items or whether the similarity of the items should be varied in terms of their distinctiveness. Therefore, we also used the model to determine whether our stimulus manipulation

would be best characterized by a change in item distinctiveness of item similarity and whether the nature of the stimuli affected the SDs that were observed.

According to all of the best fitting models, there was variability in the amount of the carryover between stimulus conditions such that more carryover occurred in the landscape condition than in the object condition. In addition, the simulations indicated that the best way to model the difference in the degree to which the objects and landscapes overlap in terms of their features was to generate items stochastically from a categorical prototype rather than by varying the base rate distribution from which the features were drawn. In fact, models in which the distinctiveness of the features was varied between stimulus conditions produced the worst fits and at times led to misleading interpretations of the data.

It is interesting to note that this may be the first finding to indicate that vigilance or attentional control during testing is influenced by the nature of the test stimuli. There are, of course, several models that assume that the nature of stimuli affect the allocation of attentional resources at study (DeCarlo, 2002, 2007; Estes & Maddox, 1997; Howard, Bessette-Symons, Zhang, Y William J. Hoyer, 2006; Malmberg & Murnane, 2002). For instance, several models assume that rare words attract more attention than common words when they are studied (Malmberg & Nelson, 2003 for a review). However, the pattern of SDs that we observed cannot be explained by fluctuations in the attention during study, since the order in which items were tested was determined randomly (also Malmberg & Annis, 2011).

Our simulations also identified several important issues. First, it appears that response bias varied between stimulus conditions. This could be better accounted for by

models in which the decision parameters were free to vary between stimulus conditions. However, it is important to note that the model does a reasonable job accounting for the positive sequential dependencies that we observed without appealing to criterion shifts either between trials or between stimulus conditions. The model could also be enhanced by a more sophisticated model of encoding. The present model was the simplest one possible. It assumed that each time an item is repeated, previously unstored features from the lexical/semantic trace representing the item are accumulated in the prior episodic trace.

Finally, in developing the model, we were confronted with the distinction between item similarity and frequency similarity. Test items vary in the extent to which they are perceptually or semantically similar and they vary to the extent that they were presented similar number of times during study. Both factors will influence the correlations among test trials. This aspect of recognition memory testing distinguishes it from perceptual testing using the absolute identification procedure where only the dimension on which the stimuli are judged at test distinguish the class of items tested. This difference between recognition testing and absolute identification may be a source of the differences in the patterns in sequential dependencies observed in memory and perception studies.

Table 1. Initial parameter estimates.

Parameter	Value
a	0.64
b	0.72
c	0.7
g	0.39
r	20.44
s	0.52*
t	18
u	0.04
w	50

**Note*: The *s* parameter was fixed at 0 when it was not varied.

Table 2. Bootstrap estimates for *g a* and *s* parameters.

Free Parameter	Condition	Mean	SD
a			
	Objects	0.777	0.009
	Landscapes	0.567	0.008
g			
	Objects	0.362	0.003
	Landscapes	0.409	0.003
s			
	Objects	0.431	0.034
	Landscapes	0.495	0.018

Table 3.Bootstrap estimates for the *a* and *g* parameters.

Free Parameter	Condition	Value	SD
a			
	Objects	0.889	0.012
	Landscapes	0.639	0.004
g			
	Objects	0.421	0.004
	Landscapes	0.444	0.002

Table 4.Bootstrap estimates for the *a* and *s* parameters.

Free Parameter	Condition	Value	SD
a			
	Objects	0.861	0.008
	Landscapes	0.491	0.012
s			
	Objects	0.285	0.070
	Landscapes	0.584	0.008

Table 5.Bootstrap estimates for the *g* and *s* parameters.

Free Parameter	Condition	Value	SD
g			
	Objects	0.396	0.003
	Landscapes	0.442	0.005
s			
	Objects	0.546	0.014
	Landscapes	0.593	0.014

Table 6.Bootstrap estimates for the *a, g*, and *s* parameters.

Free Parameter	Condition	Value	SD
a			
	Objects	0.993	0.005
	Landscapes	0.609	0.026
g			
	Objects	0.453	0.002
	Landscapes	0.433	0.005
s			
	Objects	0.362	0.051
	Landscapes	0.622	0.622

Table 7. The negation of the log-likelihood multiplied by 2, G^2, degrees of freedom (*df*), Akaiki's Information Criterion (AIC), the Bayesian Information Criterion (BIC), the change in BIC (ΔBIC) from the lowest BIC obtained, and the Bayes Factor (*B*) for each model.

Free Parameters	-2LL	G^2	*df*	AIC	BIC	ΔBIC	B	*w*
a, g, s	16509.34	762.71	163	16513.34	16527.21	0.00	1.00	1.00
a, s	16540.07	793.44	164	16544.07	16551.98	24.78	0.00	0.00
a	16726.680	980.05	165	16728.68	16732.64	205.43	0.00	0.00
a, g	16759.776	1013.143	164	16763.776	16771.688	244.48	0.00	0.00
g, s	16979.83	1233.19	164	16983.83	16991.74	464.53	0.00	0.00
g	17042.069	1295.436	165	17044.069	17048.025	520.82	0.00	0.00
s	17241.946	1495.31	165	17243.95	17247.90	720.69	0.00	0.00

Table 8. The the negation of the log-likelihood, G^2, degrees of freedom (*df*), Akaiki's Information Criterion (AIC), the Bayesian Information Criterion (BIC), the change in BIC (ΔBIC) from the lowest BIC obtained, and the Bayes Factor (*B*) for each model. This table excludes the *a, g, s* model.

Free Parameters	-2LL	G^2	*df*	AIC	BIC	ΔBIC	B	w
a, s	16540.07	793.44	164	16544.07	16551.98	0.00	1.00	1.00
a	16726.680	980.05	165	16728.68	16732.64	180.65	0.00	0.00
a, g	16759.776	1013.143	164	16763.776	16771.688	219.70	0.00	0.00
g, s	16979.83	1233.19	164	16983.83	16991.74	439.75	0.00	0.00
g	17042.069	1295.436	165	17044.069	17048.025	17048.02	0.00	0.00
s	17241.946	1495.31	165	17243.95	17247.90	17247.90	0.00	0.00

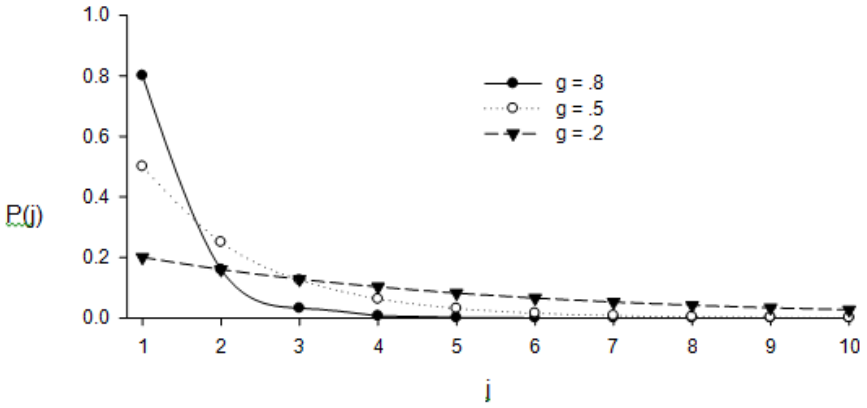

Figure 1. The probability of value *j* as a function of *j* and the geometric distribution parameter, *g*. As *g* decreases, the mean and variance of the density function increase. Thus, representations become less similar and more distinctive as *g* decreases.

Sample Landscape Items

Sample Object Items

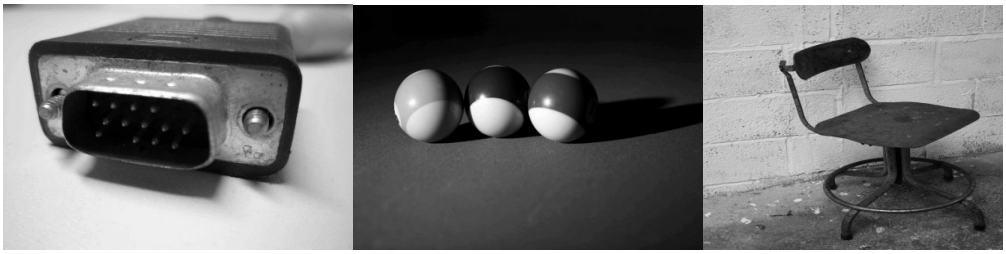

Figure 2. Sample of items presented to participants in the landscape and object condition.

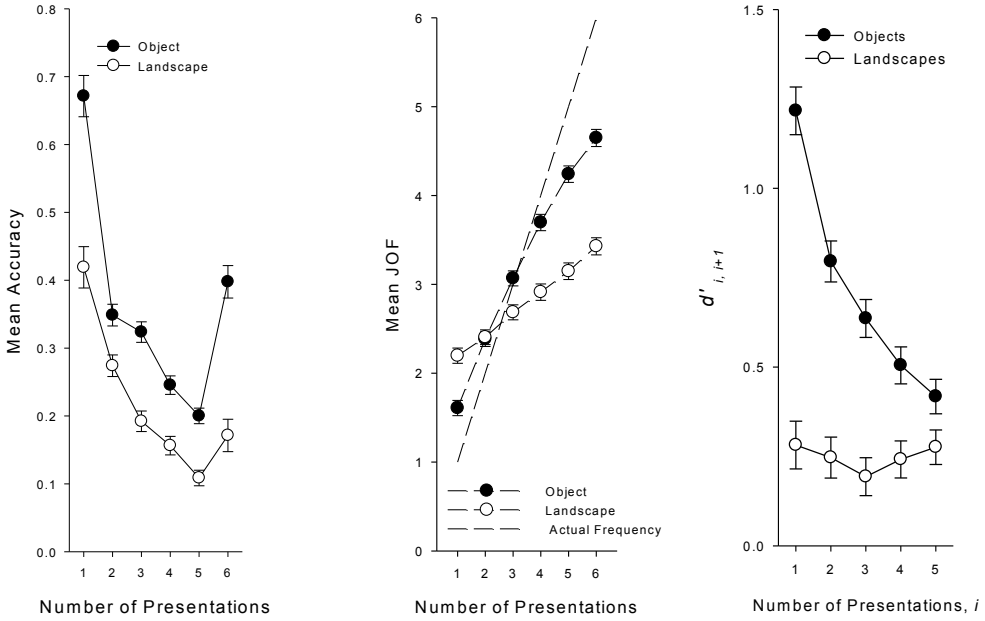

Figure 3. The left panel shows mean percent correct plotted as a function of the number of presentations. The middle panel shows the mean Judgment of Frequency as a function of the actual stimulus frequency. The right panel plots accuracy ($d'_{i, i+1}$) as a function the number of presentations, *i*.

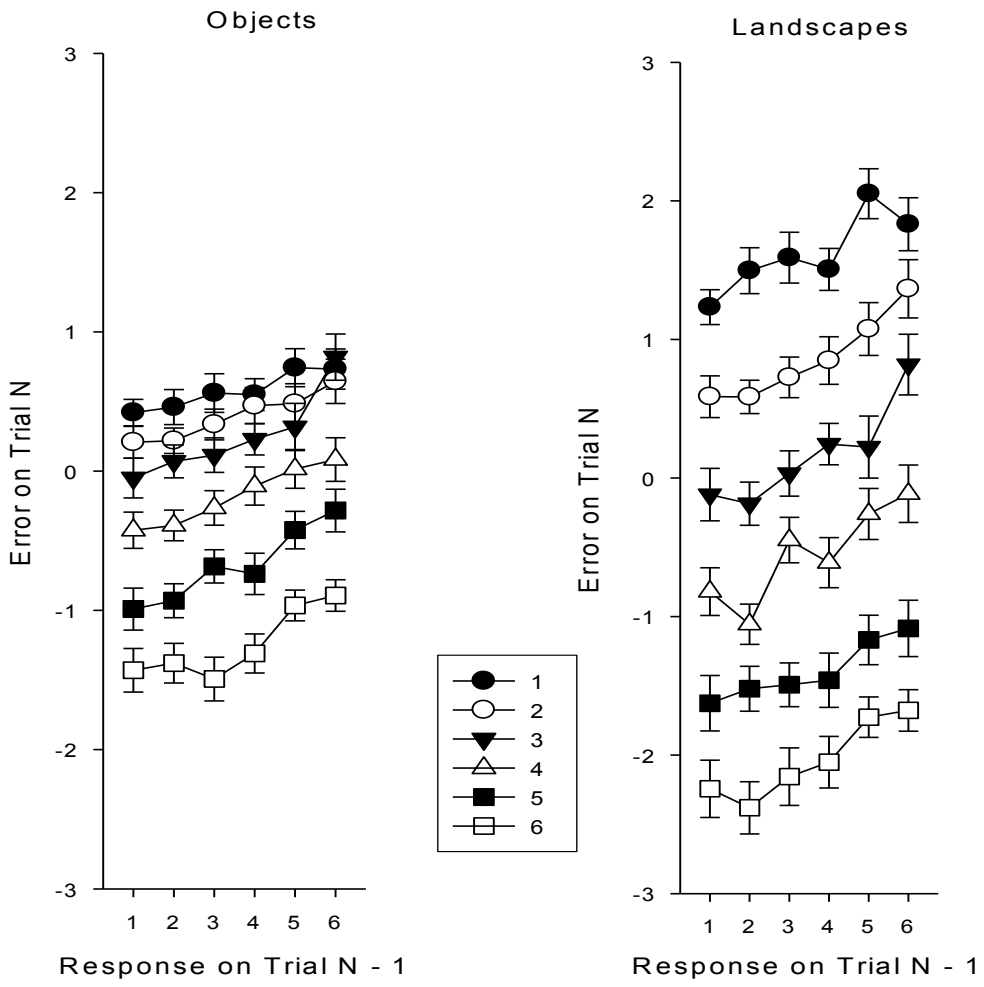

Figure 4. Mean error on trial *n* plotted as a function of the previous response and current stimulus.

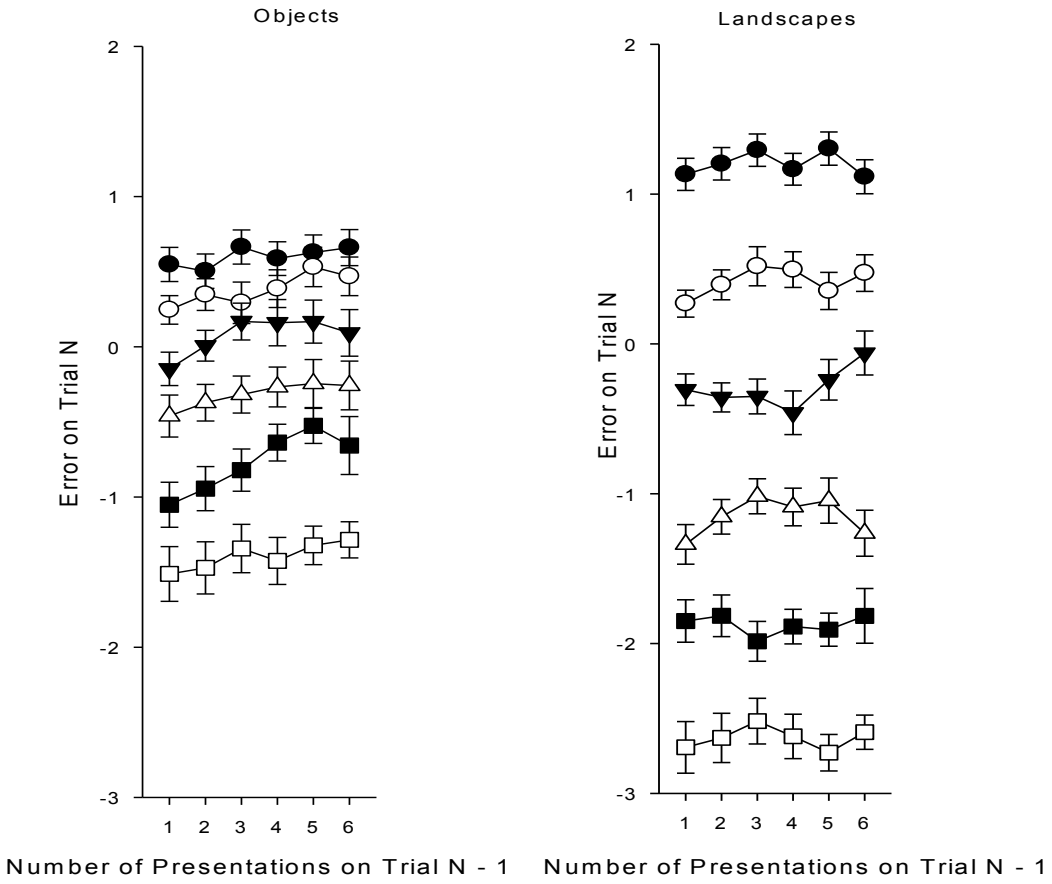

Figure 5. Mean error on trial *n* plotted as a function the previous stimulus and current stimulus values.

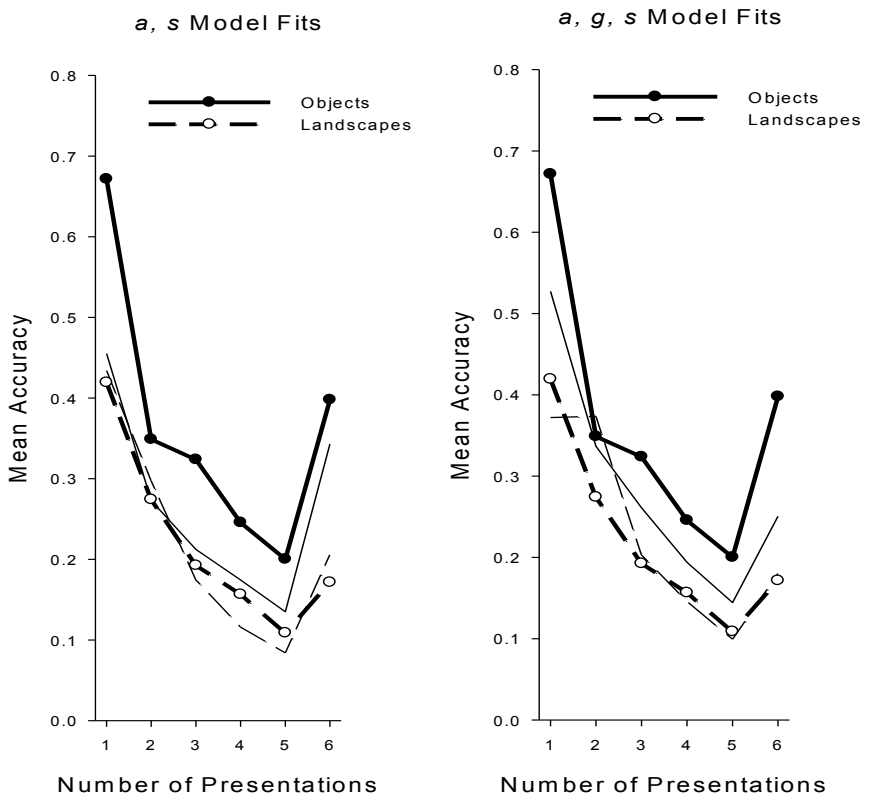

Figure 6. Model fits of the proportion correct as a function of the number of presentations.

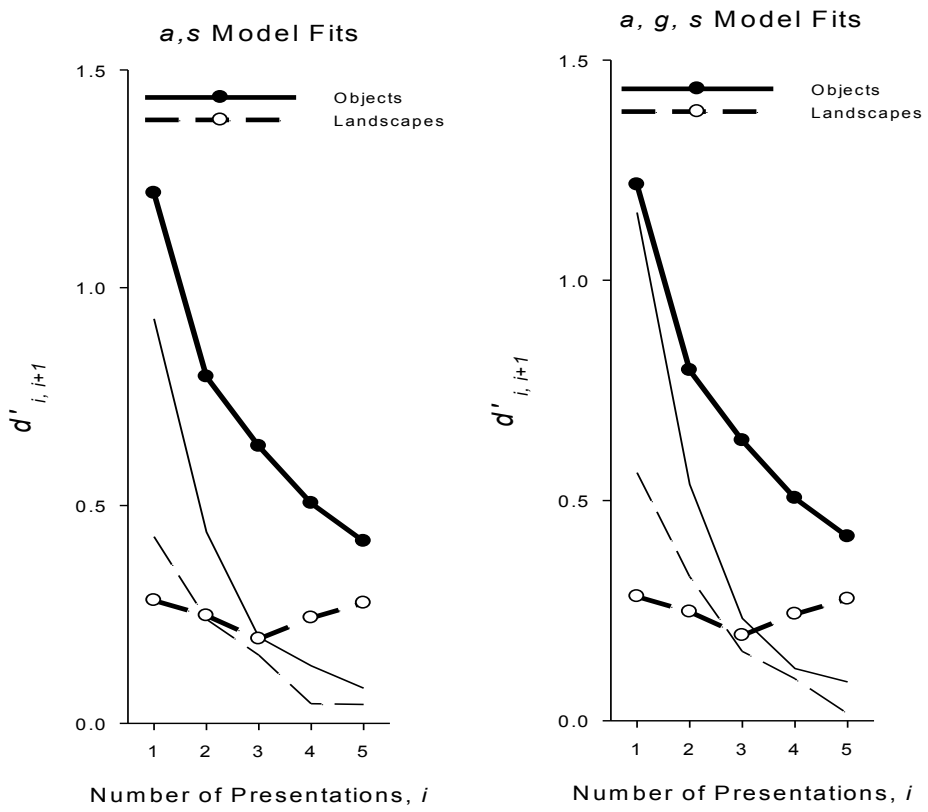

Figure 7. Model fits for accuracy ($d'_{i, i+1}$)as a function the number of presentations, *i*.

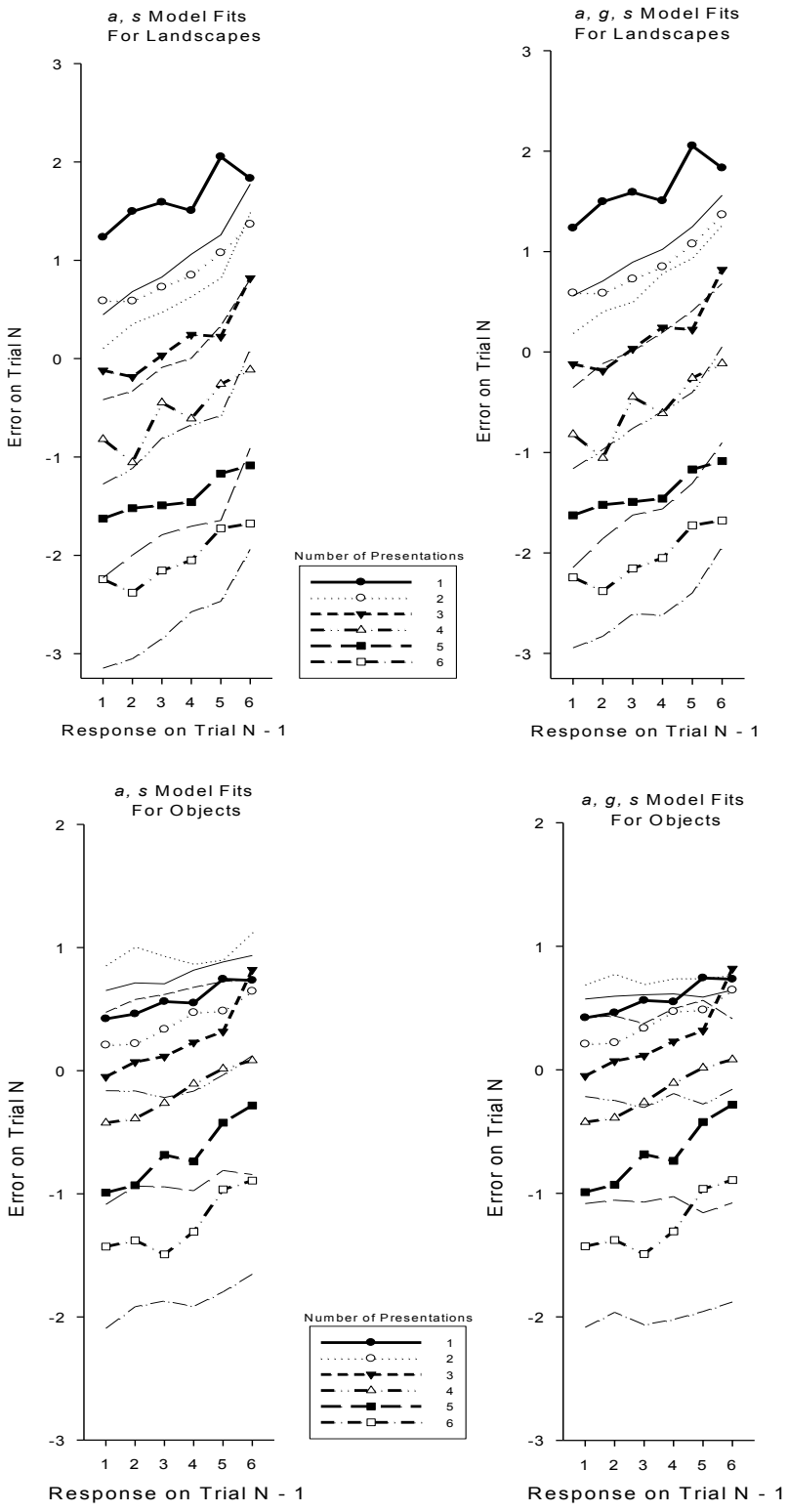

Figure 8. Model fits for error on current trial as a function of previous response.

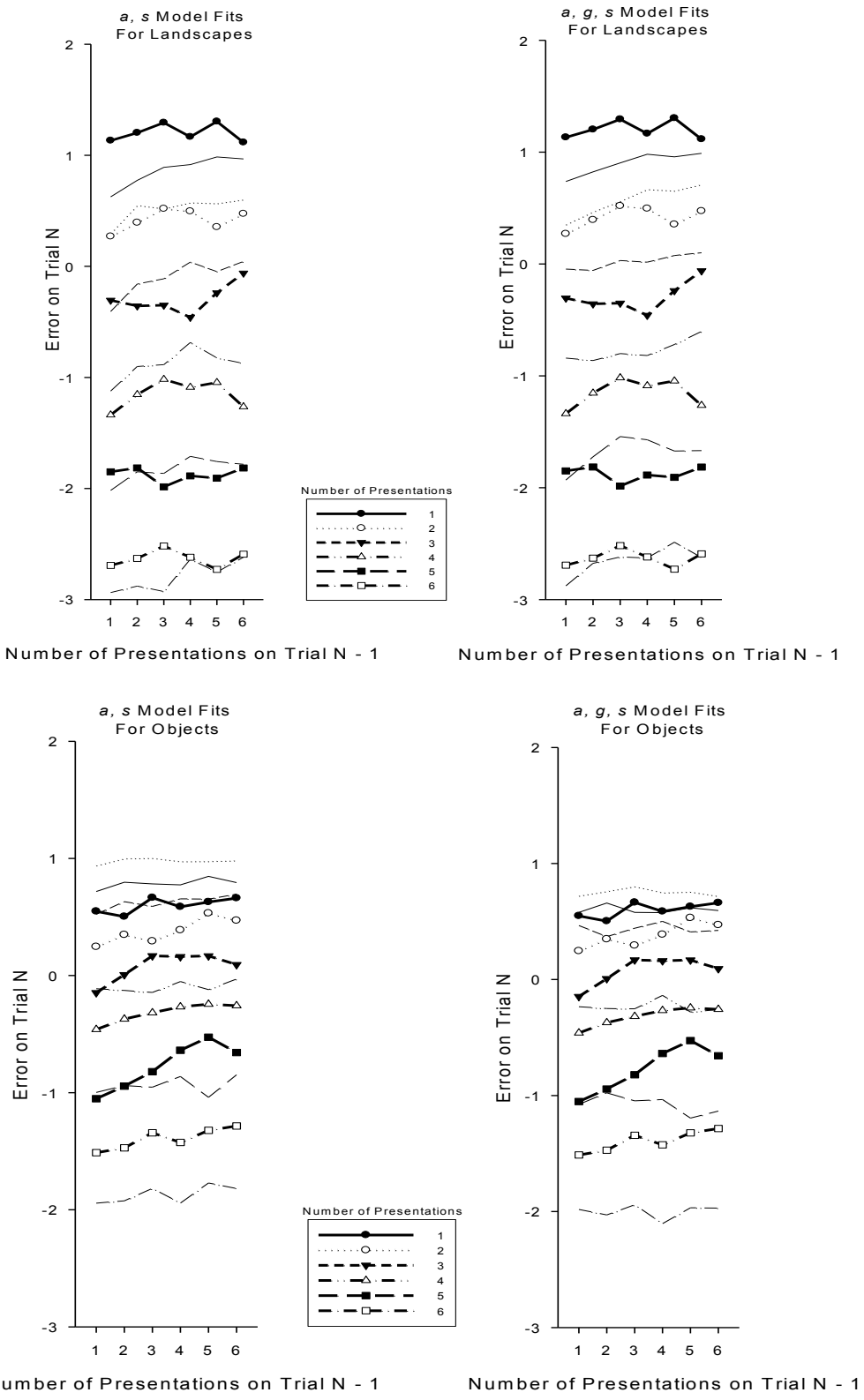

Figure 9. Model fits for error on current trial as a function of the previous number of presentations.

References

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, 74, 81–99.

Bernbach, H. A. (1970). Decision process in memory. *Psychological Review*, 74, 462–480.

Collier, G. (1954a). Intertrial association at the visual threshold as a function of intertribal Interval.*Journal of Experimental Psychology* 48(5), 330-334.

Collier, G. (1954b). Probability of response and intertrial association as functions of monocular and binocular stimulation. *Journal of Experimental Psychology* 47(2), 75-83.

Collier, G. & Verplanck, W. S. (1958). Nonindependence of successive responses at threshold as a function of interpolated stimuli. *Journal of Experimental Psychology*, 55(5), 429-437.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (No. AFCRC-TN-58-51). Bloomington, IN: Indiana University Hearing and Communications Laboratory.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace model. *Psychological Review*, 95, 528–551.

Howarth, C.I. & Bulmer, M.G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology* 8, 163-171.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review, 108*(1), 149-182.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 316-332.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and theory of signal detection. *Psychological Review*, 74, 100–109.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology, 39,* 383–395.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 319-331.

Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics, 57,* 1065–1079.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, 74, 496–504.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63,* 81–97.

Ratcliff, R. & McKoon, G. (1978). Priming in item recognition, Evidence for propositional structure of sentences. *Journal of Verbal Learning and Verbal Behavior*, 17, 403-417.

Schwartz, G., Howard, M. W., Jing, B. and Kahana, M. J. (2005). Shadows of the past: Temporal
retrieval effects in recognition memory, *Psychological Science*, 16, 898–904.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 3-11.

Tanner, T. A., Haller, R. W., & Atkinson, R. C. (1967). Signal recognition as influenced by presentation schedules. *Perception and Psychophysics*, 2, 349-358.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111.

Verplanvk, W.S., Collier, G.H., & Cotton, J.W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology, 44, 273- 282.*

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics, 9,* 73–78.