Journal of Mathematical Psychology 57 (2013) 225–236

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

A model of positive sequential dependencies in judgments of frequency

Jeffrey Annis *, Kenneth J. Malmberg

University of South Florida, United States

HIGHLIGHTS

- A model of positive sequential dependencies was developed.
- Several variants of the model were explored.
- The modeling results suggest the plausibility of the basic model.

ARTICLE INFO

Article history:
Received 8 April 2012
Received in revised form
16 May 2013
Available online 8 August 2013

Keywords:
Recognition
Sequential dependencies
Judgments of frequency
REM model
Model selection
Similarity

ABSTRACT

Positive sequential dependencies occur when the response on the current trial n is positively correlated with the response on trial $n - 1$. They are observed in a Judgment of Frequency (JOF) recognition memory task (Malmberg & Annis, 2012), and we developed a process model of them in the REM framework (Malmberg, Holden, & Shiffrin, 2004; Shiffrin & Steyvers, 1997) by assuming that features that represent the current test item in a retrieval cue carry over from the previous retrieval cue. We tested the model with data that distinguish between the number of times two given items were studied (frequency similarity) and the similarity between stimuli (item similarity), which was varied by presenting either landscape photos (high similarity), or photos of everyday objects such as shoes, cars, etc. (low similarity). Two models of item similarity were tested by assuming that the item representations share a proportion of features and that the exemplars from different stimulus classes vary in the distinctiveness or diagnosticity. A comprehensive exploration of several variants of these models directly was conducted comparing BIC and SBICR model selection statistics. The analyses establish the plausibility of the basic model of positive sequential dependencies, which assumes that differences in the similarity of the stimuli and differences in vigilance to the JOF task account for the pattern of sequential dependencies that we observed. They also indicate that different decision criteria are used to classify different stimuli on the JOF scale.

1. A model of positive sequential dependencies in judgments of frequency

Testing often assesses task performance over the course of many test trials, and many models assume the independence of the individual tests in order for them to be valid instruments for measuring or understanding. For instance, *independence* is required by many statistical tests, including maximum-likelihood analyses, analyses of variance, etc. (Anderson, 1971). Nevertheless, there are extensively documented cases in the psychological literature where the independence assumption does not hold. Absolute identification is perhaps the most well known example. In an absolute identification task, a 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 may be classified along an m-point scale by assigning an integer to each stimulus. Because the mapping of the stimulus to the response is arbitrary and the stimuli are similar, 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, violating the independence assumption (Lacouture & Marley, 1995; Mori, 1989; Mori & Ward, 1995; Ward & Lockhead, 1971).

Positive sequential dependencies (SDs) occur when the current response is positively correlated with a previous response (or stimulus), which is known as *assimilation*. If the response to the current stimulus on trial n, is S_n, and the response to the previous stimulus value on trial $n - 1$, is S_{n-1}, then assimilation is observed when the subjects' estimate of the current stimulus increases as the nominal difference between successive stimuli increases. For instance, assimilation is observed if a stimulus of category 3 is given a greater average rating following the presentation of a stimulus from category 5 than following a presentation of

* Correspondence to: Department of Psychology, University of South Florida, 4202 East Fowler Ave, PCD 4118G, Tampa, FL 33620, United States.
E-mail address: jannis@mail.usf.edu (J. Annis).

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.06.006

a stimulus from category 4. A similar pattern is often observed between the current response and the prior response (for a discussion of these different forms of assimilation see Jones, Love, & Maddox, 2006), and negative SDs or *contrast* is observed in absolute identification, typically at lags greater than 1 and only when feedback is provided (Ward & Lockhead, 1970).

SDs are also found in recognition memory tasks, including detection, confidence ratings, and judgments of frequency (Malmberg & Annis, 2012). Assimilation is observed in yes–no recognition; S_{n-1} and S_n are more likely than chance to be classified as both studied or both unstudied. Testing recognition memory via a judgment of frequency (JOF) is similar to absolute identification in that it requires a mapping of m classes of stimuli to m responses; items are studied a various number of times, and the subject responds with the number of times the word was presented at study. Although they are robust, SDs in recognition memory are not yet well understood. We therefore found it instructive in prior studies to directly compare the SD observed in JOFs to those observed in absolute identification. For instance, Malmberg and Annis (2012, Exp 5) had subjects study words from 1 to 6 times in font sizes varying at six levels. During the course of study subjects performed an absolute identification task, classifying the font size of the stimulus. Following study, recognition memory was tested via JOFs. Positive SDs between the prior response and the prior stimulus and the current response were observed for both tasks, but the patterns of negative SDs observed between recognition tasks and absolute identification were different, as noted above. Perhaps most importantly, contrast was observed in absolute identification at lags greater than 2, but contrast was only observed in the JOFs at lag 1. Thus, similar decision structures may give rise to different patterns of SDs, depending on the task. Moreover, it would not be unreasonable to assume that absolute identification and JOFs do not share the exact same mechanisms, based on the different on different patterns of SDs that are observed. But to what extent absolute identification and JOFs differ is not the goal of the current research, rather, we seek to describe a model that captures those patterns of SDs observed in recognition memory.

Here, we present the results of the initial investigation of the mechanisms underlying positive SDs in recognition memory in the form of a process-model of the JOF task conceived within the framework of the retrieving effectively from memory theory (REM, Malmberg, Holden et al., 2004; Shiffrin & Steyvers, 1997). Since, like all memory models, REM models assume that recognition is based on the outcome of an interaction between a retrieval cue and the contents of memory, the more similar the cue is to the contents of the memory, the more familiar the stimulus seems. For a JOF, we assumed that the greater the familiarity of the stimulus the greater the JOF assigned to it (Hintzman, 1988; Malmberg, Holden et al., 2004), and we conducted a series of analyses of the hypothesis that correlations in the information used to make successive JOFs are the result of a carryover of the information used to probe memory from trial to trial, perhaps as the result of lapses in attention or vigilance. 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).

2. A Model of judgments of frequency (JOFs)

REM assumes that lexical/semantic traces are represented as vectors of w geometrically distributed, feature 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: $P(V = j) = g(1-g)^{j-1}$, where $j \in \{1 \ldots \infty\}$. Most importantly, g affects the distinctiveness of the representations (Malmberg, Zeelenberg, & Shiffrin, 2004). As g increases, the mean feature value increases and the variability of the feature values increases (see Fig. 1). 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 will therefore be more distinctive compared to when the geometric distribution is defined by a relatively high g value.

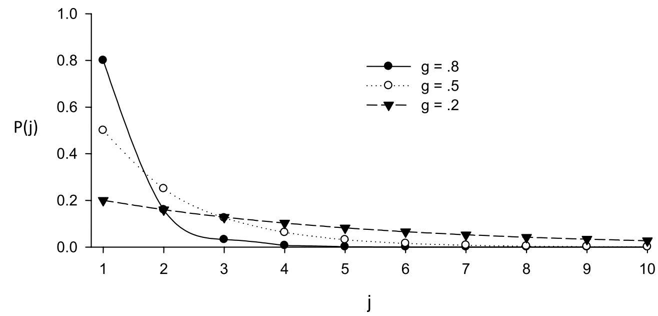

Fig. 1. The probability of value j as a function of j and the geometric distribution parameter, g. As g decreases, the mean and variance of the density function increase. Thus, representations become less similar and more distinctive as g decreases.

Repetitions. For the JOF task, items are studied one or more times on a long study list, and there are different ways to model item repetitions (Criss, Malmberg, & Shiffrin, 2011; Shiffrin & Steyvers, 1998). The simplest model assumes that each study repetition of an item results in additional storage attempts to the trace stored when the item was presented the first time (Malmberg & Shiffrin, 2005; Shiffrin & Steyvers, 1997, 1998). Therefore, the probability of storing a feature in episodic memory after studying an item r times is

$$P(\text{storage}) = 1 - (1 - u^*)^{\text{tr}} \tag{1}$$

where r is the number of times an item is repeated, and t is the number of attempts made at storing a feature on a given repetition.

We also considered a potentially richer encoding model in which a new traced is created with probability, η (Malmberg, Holden et al., 2004; Shiffrin & Steyvers, 1997, 1998). When η is 0, a repetition will never elicit the storage of a new trace in memory, which is the simple model just described, and when η is 1, a repetition will always create a new trace. When an item is repeated and an accumulation of information occurs in a trace stored on a previous presentation, unstored features (i.e. features that have a value of 0) are overwritten according to the rules described above. We assume that for all repetitions of a stimulus, only one trace, corresponding to that stimulus, accumulates features.

Retrieval. There are a number of ways to model retrieval when performing recognition (see Malmberg, 2008, for a review), and certain assumptions are necessary to achieve an appropriate level of yes–no accuracy when the stimuli are not randomly chosen (Malmberg, Holden et al., 2004; Malmberg & Xu, 2007; Xu & Malmberg, 2007). However, in the present experiment, all test items were studied, and therefore we will work with a simple model of retrieval that assumes that the test item's lexical/semantic trace serves as the retrieval cue, and it is matched in parallel to the episodic traces stored during study. For each trace, j, a likelihood ratio, λ_j, is computed:

$$\lambda_j = (1-c)^{n_{jq}} \prod_{i=1}^{\infty} \left[\frac{c + (1-c)\, g\, (1-g)^{i-1}}{g\, (1-g)^{i-1}} \right]^{n_{ijm}}, \tag{2}$$

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 value of g

used in this calculation is the same value of g used during encoding. The log odds are obtained from the average likelihood ratio for the n traces compared to the retrieval cue,

$$\phi = \ln\left(\frac{1}{n}\sum_{j=1}^{n}\lambda_j\right), \tag{3}$$

where n is the number of episodic traces stored during study, and they are next compared to a set of criteria in order to make the JOF. How these criteria are placed has a significant impact on qualitative predictions, and we will discuss different models of criterion placement in subsequent sections.[1]

Carryover. We refer to this model of assimilation as the *carryover 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). 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, as a increases, the number of trials in which carryover occurs decreases. If $a = 1$, then no carryover occurs and b is irrelevant. Similarly, when $b = 0$, the a parameter is irrelevant and the carryover process does not occur. In both of these cases, the model reverts to the Shiffrin and Steyvers (1997) model.

A tacit assumption of the model is that the subject fails to discount the information carrying over from trial-to-trial (cf. Huber et al., 2001). The failure to discount the carryover 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 Eq. (2)) and generate an odds value (see Eq. (3)). 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 2. 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 or b takes on a value of 0.

JOFs. The JOF decision mechanism proposed by Malmberg, Holden et al. (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:

$$\mathbf{C}_k = \alpha k + \beta, \tag{4}$$

where $\boldsymbol{C}_k$ is the criterion associated with the JOF k, where $k = 1 \ldots n$, α is the slope, and β is the intercept. 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.

Although biased guessing may not be a general characteristic of the JOF paradigm, when the items are generally unfamiliar, it is quite possible that participants may make use of the entire response scale rather than always responding with low JOFs since they are aware that some items were presented several times. Hence, a guessing mechanism was implemented that assumes there is a probability, γ, that the participant will randomly choose a response on the upper half of the response scale on trials where the log odds fall below the middle criterion. For example, if the response scale ranges from 1 to 6, and the log odds fall below the criterion associated with the JOF of 3, there is a probability, γ, that the participant will randomly choose, according to a uniform distribution, 4, 5 or 6. When γ is 0, no guessing occurs.

Modeling stimulus similarity. In prior JOF studies, the similarity of the stimuli has been shown to be critical (Hintzman, Curran, & Oppy, 1992; Malmberg, Holden et al., 2004). In our model of recognition memory, increasing the similarity of the items decreases their discriminability, and it would be impossible to distinguish between the items presented one time versus 6 times at the upper limit. We refer to this as *item similarity*.[2] Hence, we expected that a manipulation of item similarity would result in decreased JOF accuracy and increases in SDs, since the model predicts correlations among adjacent responses during testing to the extent that their episodic representations and retrieval cues are comprised of similar sets of features. Independent of item similarity is the dimension on which the JOFs are made, which is the unidimensional familiarity value, ϕ, obtained from the global-matching retrieval process (Eq. (2)). Each repetition condition produces a distribution over ϕ, and the more similar the number of times test items are studied, the greater the categorical prototype of their respective distributions. We refer to this dimension as *frequency similarity*. Frequency similarity is related to SDs to the extent that greater frequency similarity between adjacent test items leads to a greater correlation in the adjacent responses at test.

Here, we consider two models of item similarity. The first model that we consider 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 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 Fig. 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.

[1] For binary old–new recognition, when the log odds are greater than 0, an "old" response is given. Accordingly, the average hit rates are greater and the average false-alarm rates are lower for items generated from a geometric distribution with a relatively low g value and for items repeated relatively frequently at study.

[2] 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 absolute identification, the effect of similarity is twofold: Increases in similarity lead to both a decrease in accuracy and an increase in positive SDs (McGill, 1957; Mori, 1989; Ward & Lockhead, 1971). 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. 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.

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

3. Experiment

The carry over model creates positive sequential dependencies in recognition memory testing because the information on which a decision is made from trial to trial is correlated. However, it is unclear whether the carryover model can provide a qualitatively accurate account for assimilation among JOF responses. To assess the carryover 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. Fig. 2 shows a sample of the items presented. To manipulate frequency similarity, the items in both conditions were presented from 1 to 6 times during study, and JOFs were collected to test recognition memory. We speculated 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.

4. Method

4.1. Subjects

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

4.2. 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. Because we were interested in comparing the patterns of SDs in absolute identification to a recognition task, we did not use distractor items. Each test list consisted of the 60 images presented at study. 55 subjects completed the object condition, and 55 subjects completed the landscape condition.

4.3. Procedure

Subjects studied four lists of images and performed a math task after each. The math task consisted of mentally adding digits for 30 s. 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.

5. Results

Proportion of correct responses. A 2 (stimulus type: objects vs. landscapes) $\times 6$ (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. Panel E of Fig. 3 shows the proportion of correct responses was greater in the object condition than in the landscape condition, $F(1, 108) = 122.57$, $MSE = 0.03$, $p < 0.0005$, $\eta_p^2 = 0.53$. There was a main effect of the number of presentations on accuracy, $F(5, 540) = 104.10$, $MSE = 0.161$, $p < 0.0005$, $\eta_p^2 = 0.49$, and the stimulus type interacted with the number of presentations, $F(5, 540) = 8.19$, $MSE = 0.02$, $p < 0.0005$, $\eta_p^2 = 0.07$. To investigate the interaction, a trend analysis was conducted. The linear trend for the object condition was significant, $F(1, 54) = 56.74$, $MSE = 0.05$, $p < 0.0005$, $\eta_p^2 = 0.51$, as well as in the landscape condition, $F(1, 54) = 81.37$, $MSE = 0.03$, $p < 0.0005$, $\eta_p^2 = 0.60$. The quadratic trends for the object condition, $F(1, 54) = 138.47$, $MSE = 0.05$, $p < 0.0005$, $\eta_p^2 = 0.72$, and landscape condition were also significant, $F(1, 54) = 27.12$, $MSE = 0.03$, $p < 0.0005$, $\eta_p^2 = 0.33$. As the number of presentations increases, the

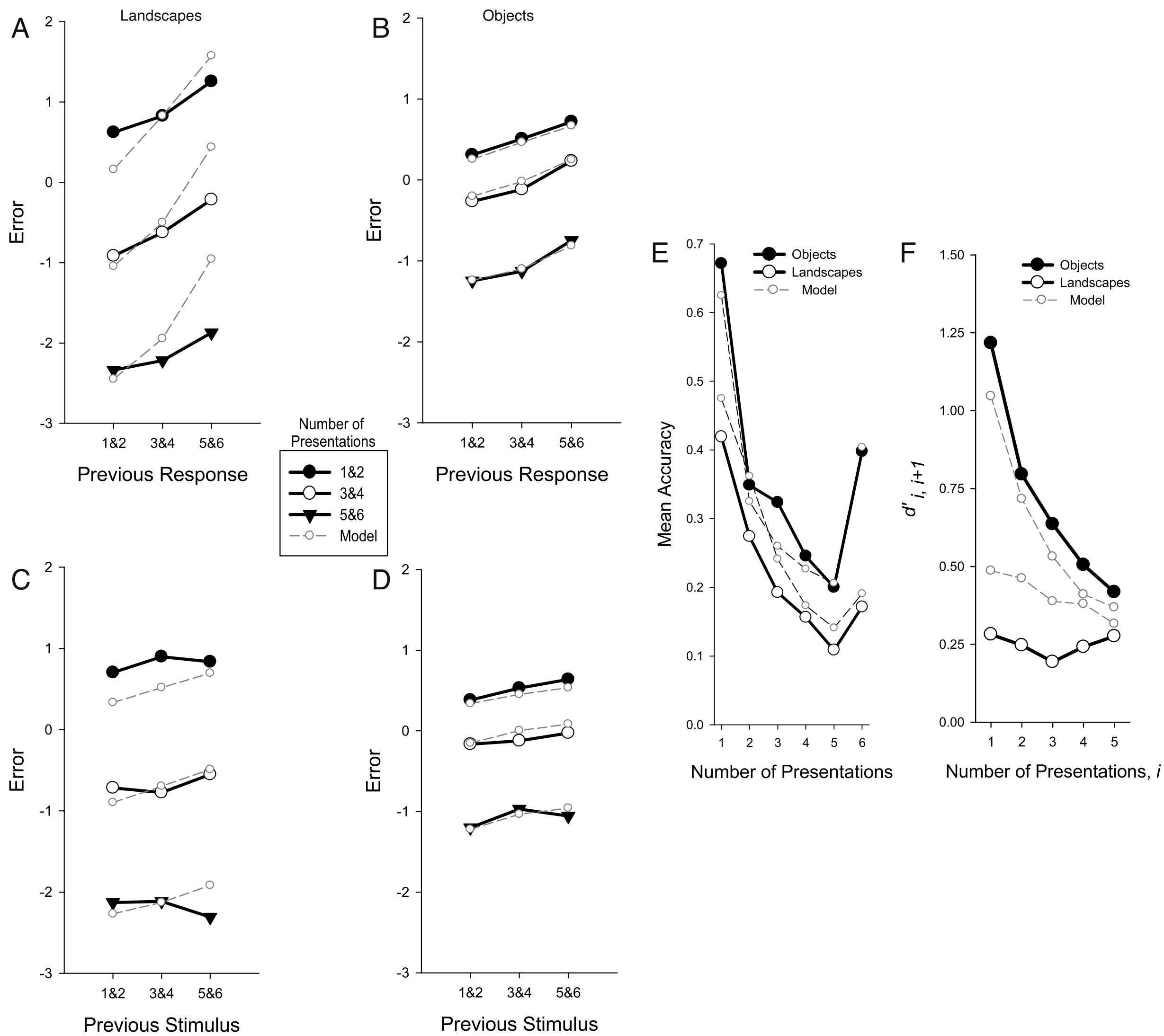

Fig. 3. Panels A, and B show the model fits for the error on current trial as a function of the previous response, while panels C, and D show the error on the current trial as a function of the previous stimulus. Panel E shows the model fits of the proportion correct as a function of the number of presentations. Panel F shows the model fits for accuracy ($d'_{i,i+1}$) as a function the number of presentations, i. The model parameters varied were g, s, and a. The parameter estimates for objects were $g = 0.46$, $s = 0$, and $a = 0.85$. The parameter estimates for landscapes were $g = 0.51$, $s = 0.54$, $a = 0.50$. Parameters values held constant: $b = 0.8$, $c = 0.75$, $t = 18$, $u = 0.01$, $w = 50$, $\alpha = 4.5$, $\beta = -4.5$, $\eta = 0$, $\gamma = 0$.

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 trough for middle range of frequencies was shallower in the landscape condition than in the object condition.

JOF accuracy. To measure JOF accuracy, $d'_{i,i+1}$ was calculated for each stimulus i and $i+1$ (Luce, Nosofsky, Green, & Smith, 1982) in order to measure the ability of the subject to discriminate items that are adjacent to each other on the frequency dimension. For instance, $d'_{2,3}$ is a measure of the ability of the subject to discriminate between items presented two versus three times. Panel F of Fig. 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) ×5 (number of presentations i) omnibus ANOVA revealed JOFs were more accurate for objects than landscapes, $F(1, 108) = 101.30$, $MSE = 0.30, p < 0.0005, \eta_p^2 = 0.48$, and while there was a main effect of the number of presentations i, $F(4, 432) = 21.08$, $MSE = 0.14, p < 0.0005, \eta_p^2 = 0.16$, $d'_{i,i+1}$ more steeply declined with increases in the number of presentations in the object condition than in the landscape condition, $F(4, 432) = 18.29$, $MSE = 0.14$, $p < 0.0005, \eta_p^2 = 0.15$. Comparing Panel E to Panel F of Fig. 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. Panels A and B of Fig. 3 plot the mean error on trial n as a function of the current stimulus and the prior response. In order to ensure that there would be enough data for the analysis, we binned the responses 1 and 2, 3 and 4, and 5 and 6. One subject from the object condition was excluded from the following analysis because they did not make every type of response for each current stimulus and prior response combination. A 2 (stimulus type) ×3 (current stimulus) ×3 (previous response) omnibus ANOVA was conducted. There were main effects of the stimulus type, $F(1, 107) = 17.59$, $MSE = 42.80, p < 0.0005, \eta_p^2 = 0.14$, and the current stimulus, $F(2, 214) = 1140.79$, $MSE = 0.38, p < 0.0005$, $\eta_p^2 = 0.914$; 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, $F(2, 214) = 180.53$, $MSE = 0.13, p < 0.0005, \eta_p^2 = 0.628$;

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, $F(2, 214) = 121.82$, $MSE = 0.38$, $p < 0.0005$, $\eta_p^2 = 0.53$, such that there was a greater underestimation of large stimulus values in the landscape condition than in the object condition. There was no previous response by stimulus type interaction, $F(2, 214) = 2.61$, $MSE = 0.27$, $p = 0.076$. There were no other significant interactions.

Panels C and D of Fig. 3 plot the mean error on trial n as a function of the current and prior stimulus. A 2 (stimulus type) ×3 (current stimulus) ×3 (previous stimulus) omnibus ANOVA was conducted. There was a main effect of stimulus type, $F(1, 108) = 17.74$, $MSE = 2.98$, $p < 0.0005$, $\eta_p^2 = 0.14$, such that the mean error in the object condition, ($M = -0.24$), was higher than in the landscape condition, ($M = -0.72$). There were also main effects of the number of presentations, $F(2, 216) = 1153.89$, $MSE = 0.38$, $p < 0.0005$, $\eta_p^2 = 0.914$, and the prior stimulus, $F(2, 216) = 13.94$, $MSE = 0.09$, $p < 0.0005$, $\eta_p^2 = 0.11$. There was a significant previous stimulus by stimulus type interaction, $F(2, 216) = 4.80$, $MSE = 0.09$, $p < 0.01$, $\eta_p^2 = 0.04$, current stimulus by previous stimulus interaction, $F(4, 432) = 9.90$, $MSE = 0.08$, $p < 0.005$, $\eta_p^2 = 0.08$ and a current stimulus by previous stimulus by stimulus type interaction, $F(4, 432) = 4.15$, $MSE = 0.08$, $p < 0.01$, $\eta_p^2 = 0.04$. However, these interactions are qualified below.

The previous response and the previous stimulus are confounded (Jones et al., 2006), and one simple way to decouple the previous stimulus from the previous response is to hold the previous stimulus constant and only let the previous response vary. We binned the data as described above and conducted a 2 (stimulus type) ×3 (current stimulus) ×3 (previous stimulus) ×3 (previous response) omnibus ANOVA in order to account for these effects. We again found no main effect of the stimulus type, $F(1, 25) = 3.99$, $MSE = 3.99$, $p = 0.057$, $\eta_p^2 = 0.14$, a main effect of the current stimulus, $F(2, 50) = 257.47$, $MSE = 5.50$, $p < 0.0005$, $\eta_p^2 = 0.91$, a main effect of the previous response, $F(2, 50) = 34.22$, $MSE = 0.79$, $p < 0.0005$, and no main effect of the previous stimulus, $F < 1$. There was a previous stimulus by current stimulus interaction, $F(4, 100) = 3.13$, $MSE = 1.61$, $\eta_p^2 = 0.11$, such that as the previous and current stimulus value increased, the error on the current trial became more negative. Thus, when the effects of the previous stimulus and response were considered independently, negative SDs away from the immediately previous stimulus were observed.

6. Modeling results

We began with six models, identified by all combinations of the a, g, and s parameters. The a parameter governs the tendency to carry over features from the previous retrieval cue to the next, and the g and s parameters determine the geometric distribution from which features are drawn and the overlap in features between stimuli, respectfully. The measures of absolute and relative goodness of fit obtained from the model fitting procedures are listed in Table 1. In addition to using the traditional BIC and AIC measures, we employed a criterion, known as the Schwartz Bayesian Information Criterion R (SBICR; Dudley & Haughton, 1997) that extends the commonly used BIC to situations in which one model is being fit to two independent datasets while varying the same set of parameters. The model with the largest SBICR is the winning model. The methods used to obtain these statistics are described in the Appendix. The SBICR and BIC both resulted in the exact same ranking for all of the models.

The first set of simulations reveals that the best fit (i.e., the model with the largest SBICR, and lowest BIC and AIC) was provided by the model in which all three of the parameters were free to vary between conditions (BIC = 6985.73; AIC = 6967.53; SBICR = −3497.36); Fig. 3 shows the best fit of the a, g, s model to the SDs and to the JOFs. The g, and s parameters were higher and the a parameter was lower in the landscape condition than in the object condition. This suggests that testing more similar stimuli inspired less vigilance from the subjects, and thus more carryover, and, most importantly, the most accurate models of the models considered take into account the sequential dependencies observed in JOFs.

Table 1
Lists the negation of the log-likelihood multiplied by 2, X^2, degrees of freedom (df), Akaike's Information Criterion (AIC), the Schwartz Bayesian Information Criterion R (SBICR) and the Bayesian Information Criterion (BIC). The table is sorted by BIC. Both the SBICR and BIC are in agreement.

Free parameters	$\ln L$	df	X	AIC	SBICR	BIC
$a, g, s, \alpha, \beta, c, \gamma$	−3419.85	15.00	17.31*	6853.69	−3458.57	6896.15
$a, g, s, \alpha, \beta, c$	−3432.27	17.00	42.16	6876.54	−3465.46	6912.93
$a, g, s, \alpha, \beta, \eta$	−3447.22	17.00	72.06	6906.45	−3480.42	6942.84
a, g, s, α, β	−3460.03	17.00	97.67	6930.05	−3487.69	6960.38
a, g, s	−3480.77	23.00	139.15	6967.53	−3497.36	6985.73
a, g	−3491.08	25.00	159.77	6986.16	−3502.14	6998.29
a	−3506.45	27.00	190.52	7014.90	−3511.98	7020.96
g, s	−3502.94	25.00	183.49	7009.88	−3514.00	7022.01
a, s	−3506.45	25.00	190.52	7016.90	−3517.51	7029.03
g	−3543.37	27.00	264.36	7088.74	−3548.90	7094.81
s	−3745.11	27.00	667.84	7492.22	−3750.64	7498.29

Note: The SBICR was based on 1595 data points from the landscape condition and 1586 data points from the object condition. The BIC was based on 3181 data points.
* Indicates $p > 0.05$.

These statistics are only useful for relative comparisons, and although the SDs are described by these simple models in both stimulus conditions, the eye-ball method is sufficient for determining that even the best fitting model is inadequate to account for the effect of stimulus type on the accuracy of the JOFs. Panels E and F of Fig. 3 show the best fit of the a, g, s model to the percent correct and the accuracy data, which is obviously quite poor; the model overestimated the bow curve and the accuracy function in the landscape condition and underestimated the bow curve and the accuracy function in the object condition. Hence, our conclusions based on this initial analysis are made provisionally and subject to revision given the need for a more complex model since it is quite possible that the carryover parameter is varying between stimulus conditions only because a more appropriate assumption has not yet been considered. Therefore, we will next consider three additional factors that might influence JOF accuracy in the different stimulus conditions.[3] One possibility is that there were different biases in the manner that the stimuli were classified. It is also possible that the objects and landscapes differed in how well they were encoded during study. Last, some of variance between conditions was undoubtedly due to random factors or noise. We consider each of these possibilities in turn in order to determine whether they help account for JOF performance.

Biases in JOF decisions. The overestimate of the bow curve and the accuracy function in the landscape condition and the underestimation in object condition may be due to the fact that the relationships between the underlying distributions of familiarity were quite different, but the same set of decision criteria were used to classify the stimuli in both stimulus conditions. Specifically, the means of the familiarity distributions are very compressed in

[3] It is important to note that because SDs are assumed to affect responses across all stimulus conditions and across all repetition conditions, that SDs do not have an overall effect on recognition accuracy. For instance, an old response to a target has the same relationship to the next response regardless of whether a target or foil is tested (Malmberg & Annis, 2012), and same holds across other dimensions of the stimuli.

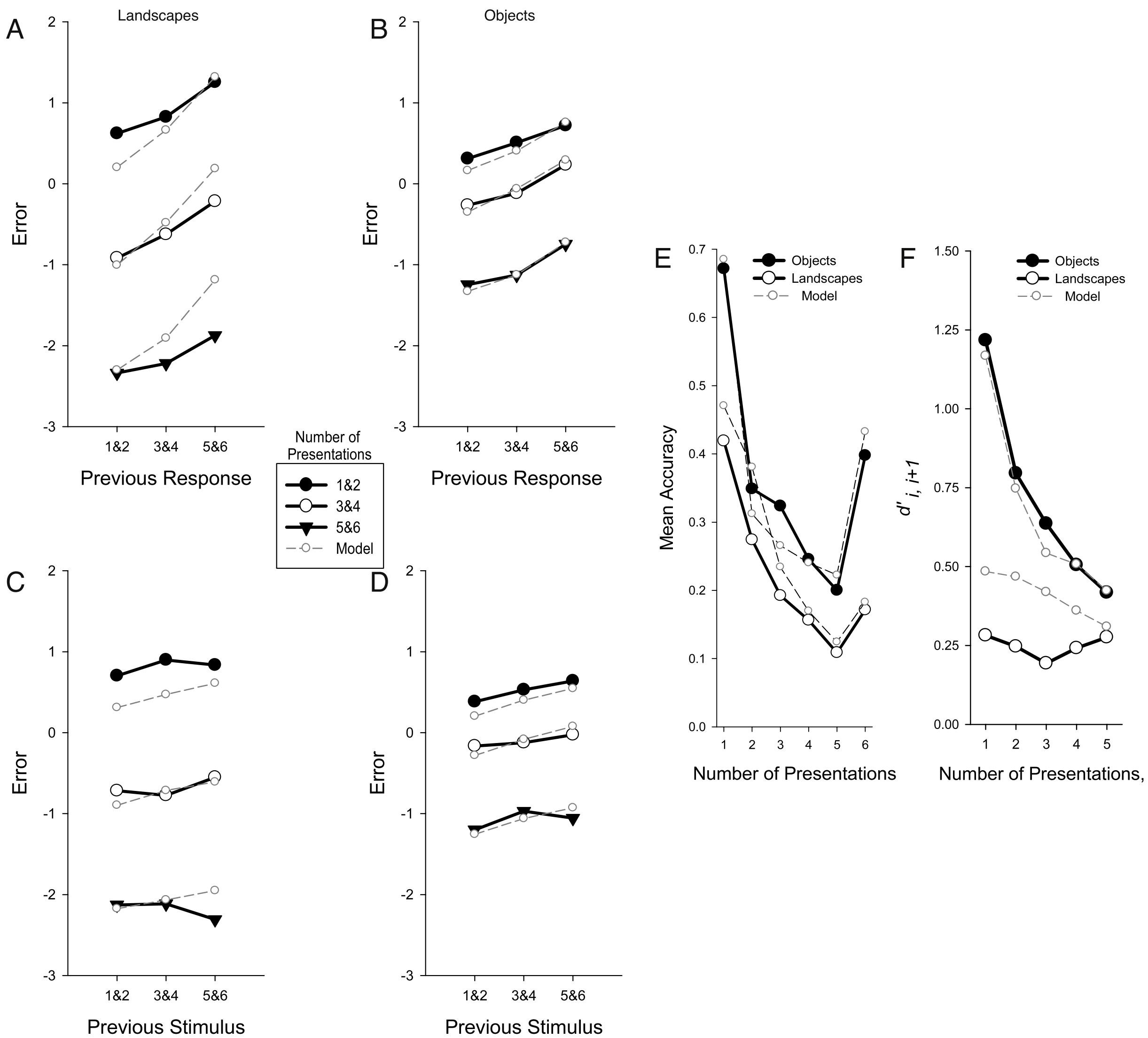

Fig. 4. Panels A, B show the model fits for the error on current trial as a function of the previous response, while panels C, and D show the error on the current trial as a function of the previous stimulus. Panel E shows the model fits of the proportion correct as a function of the number of presentations. Panel F shows the model fits for accuracy ($d'_{i,i+1}$) as a function the number of presentations, i. The model parameters varied were g, s, a, α, and β. Parameter estimates for objects: $g = 0.22$, $s = 0.09$, $a = 0.79$, $\alpha = 7.16$, $\beta = -3.39$. Parameter estimates for landscapes: $g = 0.65$, $s = 0.50$, $a = 0.67$, $\alpha = 3.45$, $\beta = -3.66$. Parameters values held constant: $b = 0.8$, $c = 0.75$, $t = 18$, $u = 0.01$, $w = 50$, $\eta = 0$, $\gamma = 0$.

the landscape condition, which is certainly one factor that affects accuracy. Therefore, it is possible that restricting the models to use the same decision criteria would lead to suboptimal JOFs in one or both stimulus conditions. For instance, if the lower criteria used to classify the objects were too conservative for the classification of the landscapes, then subjects would tend to underestimate the frequency with which landscapes were repeated.

To explore this possibility, we added the decision parameters, α and β, to the a, g, s model and allowed them to vary freely between stimulus conditions. The fits of the model are shown in Fig. 4. Table 1 shows that these modifications improved the qualitative and quantitative fits (BIC = 6960.38; AIC = 6930.05; SBICR = −3487.69). The model provided a significantly better fit to the data than the simpler a, g, s model, $X^2(2) = 41.48$, $p < 0.01$. Nevertheless, as shown in Panel F Fig. 4, the model still overestimated the accuracy in the landscape condition. It is also interesting to note that the difference in the carryover parameter observed between the two stimulus conditions decreased, compared to the simpler model, suggesting that the strain to account for difference in JOF accuracy by the simpler model was artificially being captured by a change in SDs.

Differences in encoding. Two different models of encoding were considered. The first model varied the way traces are stored in memory. For example, it is possible that traces do not always accumulate in the same trace on each repetition and instead a new trace is formed. This was modeled with the η parameter. Therefore, this model varied the a, g, s, α, β, and η parameters between conditions. The second encoding model varied the accuracy of encoding. In REM, this is governed by the c parameter. Therefore, in this model, the a, g, s, α, β, and c parameters were all free to vary between conditions.

The results of the model fitting procedure, shown in Table 1, revealed that both the models in which c and η were varied improved the fits over models that did not consider variations in encoding across stimulus conditions, $X^2(2) = 55.51, p < 0.01$ and $X^2(2) = 25.61, p < 0.01$, respectively. However, varying c was shown to improve the fit of the model (BIC = 6912.93; AIC = 6876.54; SBICR = −3465.46) more so than when η was allowed to vary (BIC = 6942.84; AIC = 6906.45; SBICR = −3480.42). The accuracy of encoding decreased from the object to the landscape condition, while the amount of new traces formed, increased from object to the landscape condition. The fits of the model in which c

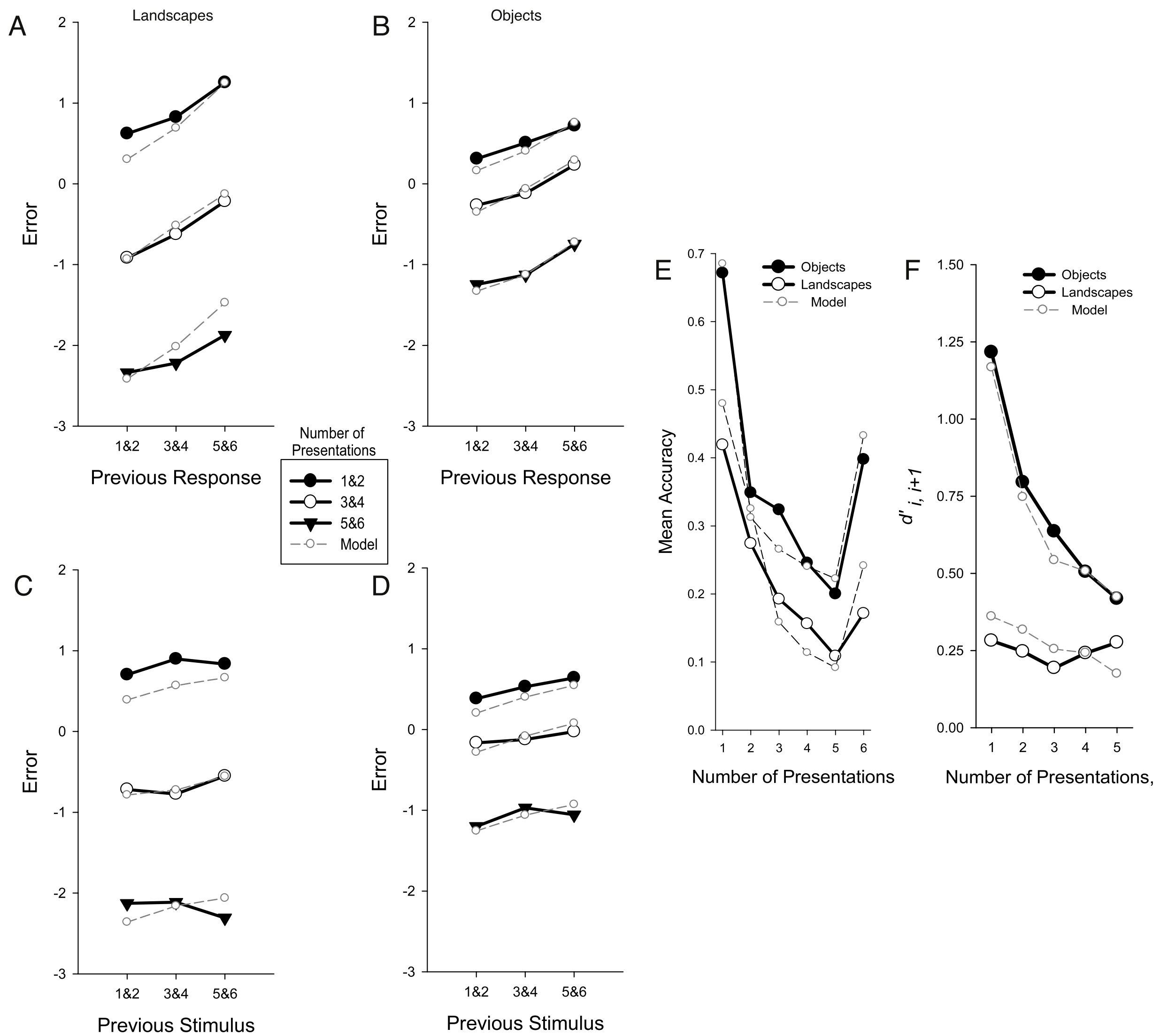

Fig. 5. Panels A, B show the model fits for the error on current trial as a function of the previous response, while panels C, and D show the error on the current trial as a function of the previous stimulus. Panel E shows the model fits of the proportion correct as a function of the number of presentations. Panel F shows the model fits for accuracy ($d'_{i,i+1}$) as a function the number of presentations, i. The parameter estimates varied were g, s, a, α, β, and c. The parameter estimates for landscapes were $g = 0.66$, $s = 0.17$, $a = 0.68$, $\alpha = 1.49$, $\beta = -3.09$, $c = 0.51$. Parameter estimates for objects (same as in Fig. 4): $g = 0.22$, $s = 0.09$, $a = 0.79$, $\alpha = 7.16$, $\beta = -3.39$. Parameters values held constant: $b = 0.8$, $t = 18$, $u = 0.01$, $w = 50$, $\eta = 0$, $\gamma = 0$.

was varied are shown in Fig. 5. Note that for the object condition, we were unable to identify parameter estimates that improved the fit of the model over the previously considered model in which encoding was not varied.

Random influences. Finally, a guessing mechanism, outlined above, was implemented, governed by the γ parameter. The fits of the model are shown in Fig. 6. Therefore, this model involved varying the following parameters: a, g, s, α, β, c, and γ. Subjects were more likely to guess when in the landscape condition than when in the object condition. This model was shown to provide the best qualitative and quantitative fit to the data. The model obtained the highest SBICR (−3484.37), lowest BIC (6896.15) and AIC (6853.69), and provided a significantly better fit than the next best model, $X^2(2) = 24.85$, $p < 0.01$. Note that in this model there is virtually no difference in the carryover parameter between the stimulus conditions, even though it was free to vary. The differences in accuracy between the stimulus conditions were better handled by changes in the guessing parameter. The model was able to quantitatively fit the object data, $X^2(22) = 3.36$, $p > 0.05$, and the landscape data, $X^2(22) = 13.94$, $p > 0.05$.

The model also fits another more subtle aspect of the data. For objects, the difference in error magnitudes between stimuli on the current trial increases with stimulus frequency. For example, the error magnitude difference between the stimuli presented once and twice is lower than the error magnitude difference between stimuli presented five and six times. This is not the case in the landscape condition. The difference in error magnitude is constant across stimulus frequencies. The model was able to capture the different trends in the landscape and object conditions. The model was not specifically designed to behave in such a manner, rather the behavior stems from a small number of reasonable assumptions about the nature of human memory and the decisional processes involved.

7. General discussion

We devised an account of sequential dependencies (SDs) in recognition testing that produces a systematic error in the signal obtained from memory. The noise is produced by a process which is referred to as carryover, since it is assumed that a certain proportion of features representing the consecutively tested items sometimes carry over from trial to trial. Carryover was implemented within the REM framework, and we tested several recognition

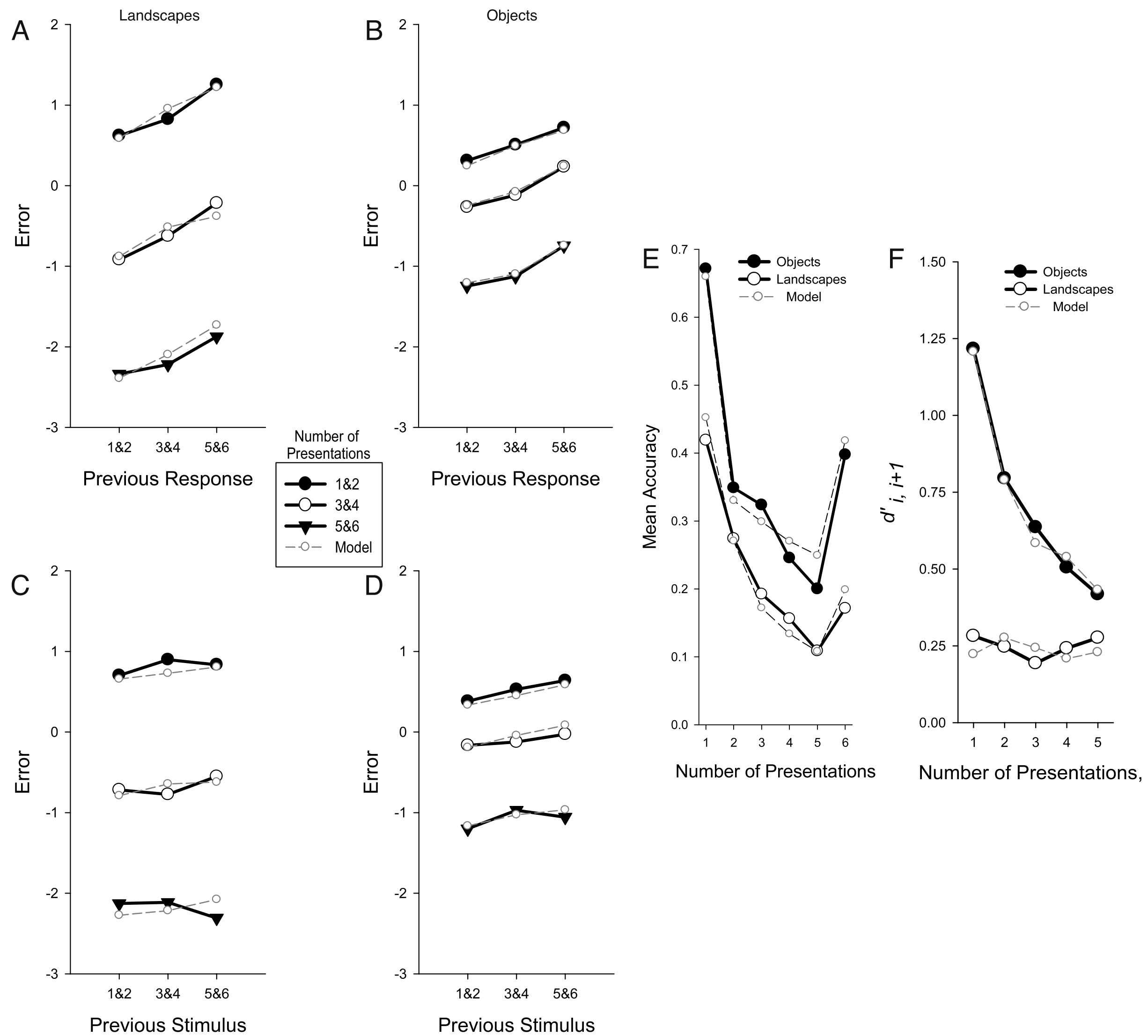

Fig. 6. Panels A, B show the model fits for the error on current trial as a function of the previous response, while panels C, and D show the error on the current trial as a function of the previous stimulus. Panel E shows the model fits of the proportion correct as a function of the number of presentations. Panel F shows the model fits for accuracy ($d'_{i,i+1}$) as a function the number of presentations, i. The parameter estimates varied were $g, s, a, \alpha, \beta, c$, and γ. The parameter estimates for landscapes were $g = 0.49, s = 0.41, a = 0.75, \alpha = 1.77, \beta = -1.78, c = 0.43, \gamma = 0.12$. Parameter estimates for objects: $g = 0.39, s = 0.00, a = 0.80, \alpha = 6.69, \beta = -4.65, c = 0.86, \gamma = 0.03$. Parameters values held constant: $b = 0.8, t = 18, u = 0.01, w = 50, \eta = 0$.

models that differed significantly in complexity with JOF data obtained from an experiment in which both stimulus similarity and item similarity were varied. Our analyses indicate that the carryover model is an adequate account of the positive SDs in JOFs, and assumptions concerning encoding and decision making differences are required to account for the effect of item similarity in the accuracy of JOFs.

Carryover in the broader context. One way to interpret the carryover model is that vigilance or attentional control is required during recognition testing (cf. Jacoby, 1991), and especially during lapses of attention, the information used to make prior recognition decisions is erroneously combined with the information used to make a new decision. Part of the evidence to support this assumption is that the fitting procedure suggested that carryover only occurred on about 20%–30% of the trials; post hoc attempts to fit the model on the assumption that carryover occurred on every trial were unsuccessful. Hence, we speculate that periodic lapses of attention introduce systematic forms of noise in recognition memory testing, and this noise deflects the present response in the direction of prior response because the subject is unaware that the signal received from memory consists of noise in the form of information about the prior occurrence of earlier items in addition to information about the prior occurrence of the current test item. The fact that positive SDs are seemingly ubiquitous in recognition testing (Malmberg & Annis, 2012), and they varied little between stimulus conditions when measured via the carryover parameter, but other parameters did, suggests within the framework of our model that lapses in attention are a relatively general characteristic of recognition testing.

Many recent models of absolute identification also assume that positive SDs arise from information that carries over from trial-to-trial (Brown et al., 2008; Petrov & Anderson, 2005; Stewart et al. 2005), and one might speculate that these models would be useful for accounting for recognition memory testing. However, the SDs found in recognition testing are different than those found in absolute identification (Malmberg & Annis, 2012), and therefore it is unclear how useful models of absolute identification would be in accounting for JOFs since those models make no predictions concerning the performance of other tasks. It is, of course, possible that positive SDs in absolute identification are also related to lapses in attention during perceptual testing.

In contrast, the framework within which we work accounts for dozens of episodic memory phenomena, and we generalized some assumptions from related REM models. For instance, in developing our model of positive SDs, we noted that there is a strong relationship between short-term recognition and long-term recognition (e.g., Nelson & Shiffrin, 2013). In fact, the carryover mechanism we proposed is inspired by short-term recognition findings that suggest features representing different items presented in a temporal sequence get combined at a cognitive level of representation without the subject's knowledge (Huber et al., 2001; Sanborn, Malmberg, & Shiffrin, 2004). Hence, our carryover model can be straightforwardly viewed as an extension of the ROUSE short-term recognition model, implemented in a REM long-term recognition framework. The carryover model is also an extension of the JOF model used to account for the interactions of normative word frequency, item similarity, and repetitions observed in recognition testing (Malmberg, Holden et al., 2004), which assumed that word-frequency was correlated with the distinctiveness of the features used to represent words (Malmberg et al., 2002; Shiffrin & Steyvers, 1997), and that item similarity is varied by manipulating the proportion of shared features among items constructed from a given base rate distribution of feature values.

There are several models that assume that the nature of stimuli affects the allocation of attentional resources during the study phase (DeCarlo, 2002, 2007; Howard, Bessette-Symons, Zhang, & Hoyer, 2006; Maddox & Estes, 1997; Malmberg & Murnane, 2002). For instance, several models assume that rare words attract more attention than common words when they are studied (see Malmberg & Nelson, 2003, for a review). In REM, rare words are distinguished from common words by lowering the g parameter, thereby making the features more distinctive (Malmberg, Zeelenberg et al., 2004). Intuitively speaking, object photographs would likely contain more distinctive features than landscape photographs on average. Indeed, this intuitive assumption is corroborated by the best fitting models involving variations in the g parameter across stimuli. On this note, it may not be overly surprising that lapses in attention would be slightly more prevalent during testing in the landscape condition than in the object condition, but the difference between the current data and previous findings is that variations in vigilance across stimuli were found at test, while this effect has been speculated to only occur at study, and of course, SDs cannot be explained by fluctuations in the attention during study, since the order in which items were tested was determined randomly (also Malmberg & Annis, 2012).

Stimulus versus item similarity. JOFs are particularly interesting because the similarity of stimuli may be varied along dimensions on which the JOF is made (i.e., number of presentations) and along dimensions that are independent of the JOF scale. 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. 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. The present experiment varied the stimuli on both dimensions by using highly similar landscape photos in one condition and dissimilar object photos in the other condition. Hence, the challenge of the model was to simultaneously account for both, and the empirical question was how item similarity would affect SDs in JOF. Although we observed a large drop in accuracy in the landscape condition, we did not observe an increase in assimilation. This was an unexpected finding insofar as SDs are only obtained when accuracy is compromised, and therefore the present finding is a challenge for any model of SDs. In absolute identification tasks, for instance, increased stimulus similarity causes accuracy to decrease and assimilation to increase (McGill, 1957; Mori, 1989; Ward & Lockhead, 1971).

Three correlates of item similarity. There were three factors besides the similarity of stimuli and variability in the amount of carryover that we found necessary to model in order to account for the accuracy of the JOFs. According to the model, there is variability in the amount of feature overlap and the number of times items were presented, and these differences influence the distribution of odds values in each condition. There is therefore no single optimal set of decision criteria for all conditions of the experiment, and it is natural to assume, therefore, that subjects vary their decision criteria when presented with these different distributions of odds values. By varying the decisional criteria, the model was shown to provide a better fit to the data than when assuming the criteria stay constant across stimulus conditions. In addition, the best fitting model indicated that encoding was noisier in the landscape condition than in the object condition. This is an interesting result, as there is a large body of evidence that shows the amount of attention a stimulus receives is positively related to the strength of the encoded trace and stimuli that possess more distinctive features receive more attention on average (see Malmberg & Nelson, 2003, for a review), but there does not appear to be such differences in vigilance associated with recognition testing insofar as the positive SDs were about the same in the landscape and object conditions. Last, we considered the influence of guessing. We assumed that subjects make use of the entire response scale. For example, the subject knows that they were presented with a number of stimuli that were presented six times. However, if the odds values are consistently low, as was the case in the landscape condition, subjects may systematically respond with a higher JOF than would otherwise be elicited to ensure that the entire response scale is being utilized.

Negative sequential dependencies. We also observed that the error on the current trial became more negative as the previous and current stimulus value increased (also Malmberg & Annis, 2012). On each trial, the JOF given by the subject corresponds to the greatest criteria exceeded by the log odds value, and the log-odds range from negative infinity to positive infinity. Therefore, for every set of criteria that exists, one must consider the special case of a log-odds value falling below the lowest criterion. In fact, the log-odds associated with an item may fall below zero, which suggests that subjects did not recognize its prior occurrence. When the subject does not have evidence that the item was studied, he may simply map these log-odds values to the lowest criterion. Because most stimuli in the current experiment were presented more than one time, this type of decisional mechanism would lead to a systematic underestimation of stimuli that were presented more than once. That is, the error on the current trial would become more negative as the stimulus value increased, which is a negative sequential dependency at the lag which we observed here and elsewhere (Malmberg & Annis, 2012).

8. Conclusions

Although it is not without some detractors (Treisman & Williams, 1984), a consensus is emerging that positive sequential dependencies in cognitive testing are attributable to the persistence of information across trials. Here we presented a carryover model of positive sequential dependencies in recognition testing that accounts for the positive SDs found in JOFs during recognition memory testing. Our model is different from those found in the absolute identification literature in that carryover is only thought to occur on a small subset of trials, whereas carryover occurs on all trials in models of absolute identification (Brown et al., 2008; Petrov & Anderson, 2005; Stewart et al. 2005). We speculate within the present framework that carryover is the result of lapses of attention during recognition testing.

Appendix. Fitting methods

When choosing the constant parameter values, we simply chose "default REM" starting values that allowed the model to behave in a reasonable manner based on dozens of other simulations that we have reported over the years and qualitatively fit both conditions equally well. When assessing the role of more complex assumptions, we borrowed the best fitting parameter values from the prior simulations in order to determine how the additional assumptions may tune the model's fit to the data better. It is possible that other constant values could have been chosen. After defining the constant and initial parameter values, the Particle Swarm Optimization algorithm (PSO; Trelea, 2003) was used to estimate the maximum of the likelihood function. The PSO technique was used because of the high-dimensionality of the parameter space and the large number of local minima. PSO is a suitable algorithm to use for such instances (Trelea, 2003). If the error gradient did not change by a tolerance of 1×10^5 for 20 iterations, the PSO algorithm would stop and report the parameter estimates. To ensure a stopping point for the PSO procedure, a maximum of 100 objective function evaluations were allowed. On each iteration of the PSO algorithm, 30 "particles", each representing a set of parameter values, were evaluated. Thus, the PSO algorithm evaluated 3000 different sets of parameter values in total. Each evaluation consisted of 100 simulated subjects who completed 10 lists.

In order to evaluate the goodness-of-fit, the Likelihood Ratio Test (LRT) was used according to Eq. (A.1).

$$G^2 = -2\ln L_{\text{specific}} - (-2\ln L_{\text{saturated}}). \tag{A.1}$$

The G^2 statistic 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 $L_{\text{saturated}}$ is the log-likelihood of the saturated model. The saturated model contains a parameter for each data point. The log-likelihood for the saturated model is calculated by Eq. (A.2).

$$\ln L\left(x_i|u, \sigma^2\right) = \ln\left(\left[\frac{1}{\sum_i (x_i - u)^2}\right]^{\frac{n}{2}} \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{n}{2}\right)\right), \tag{A.2}$$

where x_i is the ith data point, u is the mean of the data. The specific model's log-likelihood is calculated similarly, but instead of using the mean of the data as u, the mean of the model's prediction is used.

An extension of the Schwartz Bayesian Information Criterion (SBIC; Schwartz, 1978), known as Schwartz Bayesian Information Criterion R (SBICR; Dudley & Haughton, 1997), was used in the model selection process. This criterion is used when it is necessary to select a single model given multiple independent datasets for which the parameters may vary. For example, if there are 2 independent datasets and a model was fit to both datasets by varying the same set of parameters, the SBICR would be an appropriate criterion to select the best model. While it is possible to calculate the BIC for a single dataset, this may result in different winning models for each dataset. The SBICR, therefore, offers an elegant solution to this type of problem. The SBICR is given by Eq. (A.3).

$$\text{SBICR} = \sum_{i=1}^{K} \ln L_i - \frac{k}{2}\sum_{i=1}^{K} \log(n_i) + \frac{K}{2} k \ln(2\pi), \tag{A.3}$$

where the first term represents the overall log-likelihood across all K independent datasets. The second term is a penalty term that takes into account the sample size n_i for each dataset, i, and the number of parameters, k, in the model. Finally, the third term, known as a *bonus*, is added to the first two terms. The model with the largest SBICR is selected.

References

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.
Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*(2), 396–425.
Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(4), 316–326.
DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721.
DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 18–33.
Dudley, R. M., & Haughton, D. (1997). Information criteria for multiple data sets and restricted parameters. *Statistica Sinica*, *7*, 265–284.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.
Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 667–680.
Howard, M. W., Bessette-Symons, B., Zhang, Y., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: evidence from modeling and receiver operating characteristic curves. *Psychology and Aging*, *21*(1), 96–106.
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.
Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.
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.
Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology*, *39*, 383–395.
Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A. F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics*, *32*(5), 397–408.
Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 539–559.
McGill, W. J. (1957). Serial effects in auditory threshold judgments. *Journal of Experimental Psychology*, *53*, 297–303.
Malmberg, K. J. (2008). Recognition memory: a review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384.
Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition testing. *Journal of Experimental Psychology: General*, *141*(2), 233–259.
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.
Malmberg, K. J., & Murnane, K. (2002). List composition and the word- frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 616–630.
Malmberg, K. J., & Nelson, T. O. (2003). The word-frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition*, *31*(1), 35–43.
Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 322–336.
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature-frequency effects in recognition memory. *Memory & Cognition*, *30*(4), 607–613.
Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*, 540–549.
Malmberg, K. J., & Xu, J. (2007). On the flexibility and on the fallibility of associative memory. *Memory & Cognition*, *35*(3), 545–556.
Mori, S. (1989). Effects of stimulus information and number of stimuli on sequential dependencies in absolute identification. *Canadian Journal of Experimental Psychology*, *52*(2), 72–83.
Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics*, *57*, 1065–1079.
Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394.
Petrov, A., & Anderson, J. R. (2005). The dynamics of scaling: a memory-based anchor model of category rating and absolute identification. *Psychological Review*, *112*(2), 383–416.
Sanborn, A. N., Malmberg, K. J., & Shiffrin, R. M. (2004). High-level masking effects on perceptual identification. *Vision Research*, *44*(12), 1427–1436.
Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.
Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford, England: Oxford University Press.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881–911.
Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*, 68–111.
Trelea, I. C. (2003). The particle swarm optimization algorithm: convergence analysis and parameter selection. *Information Processing Letters*, *85*(6), 317–325.
Ward, L. M., & Lockhead, G. R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, *84*(1), 27–34.
Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics*, *9*, 73–78.
Xu, J., & Malmberg, K. J. (2007). Moldeing the effects of verbal-and nonverbal-pair strength on associative recognition. *Memory & Cognition*, *35*(3), 526–544.