Recognition and position information
in working memory for visual textures

YUKO YOTSUMOTO
Brandeis University, Waltham, Massachusetts

MICHAEL J. KAHANA
University of Pennsylvania, Philadelphia, Pennsylvania

AND

CHRIS MCLAUGHLIN AND ROBERT SEKULER
Brandeis University, Waltham, Massachusetts

In three experiments, we examined connections between item-recognition memory and memory for item-position information. With sequences of compound gratings as study and probe items, subjects made either item-position judgments (Experiments 1 and 2), by identifying the serial position of the study item that matched the probe, or recognition judgments (Experiment 3), by judging whether the probe had or had not been presented in the study series. Integrating a summed-similarity account of recognition into a signal detection framework shows that the variance of summed similarities on lure trials (probe not present in the study series) exceeds the variance on target trials (probe present in the study series). This prediction is borne out by the empirical *z*ROC functions, all of which had slopes that were greater than 1. Additionally, about 25% of correct recognitions were accompanied by incorrect item position identification. Misidentifications of item position arose from two sources—structural similarity and positional similarity—which combined in an approximately additive fashion.

Working within a framework of exemplar-similarity models of memory (see, e.g., Estes, 1994; Kahana & Sekuler, 2002; Kahana, Zhou, Geller, & Sekuler, 2007; Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky & Kantner, 2006), we used sinusoidal luminance gratings as stimuli in a modified Sternberg (1966) recognition task. The metric properties of the grating stimuli were exploited to test a novel prediction generated by combining the exemplar-similarity approach with an explicit, signal detection account of decision making (Wickens, 2002).

In exemplar-similarity models of recognition memory, it is assumed that a summed-similarity computation is a basic component of subjects' recognition judgment. This computation sums—over all study items—the **p**'s similarity to each of the study items. According to the model, when this sum reaches or exceeds some critical value, the subject will say "yes," judging that the **p** had been among the *n* study items that had just been seen. Following convention, we will use the term *target* (*T*) to designate trials on which **p** replicated a study item and the term *lure* (*L*) to designate trials on which **p** did not replicate any of the study items. On average, the value of summed similarity on *T* trials will exceed that on *L* trials, which means that *P*(yes) responses on *T* trials will be higher than those on *L* trials. The nature of the elements entering into the computation will also

tend to produce a systematic difference in the variances of summed-similarity values on *T* and on *L* trials, which leads to an unexpected prediction for the slope of *z*-transformed receiver operating characteristics(*z*ROCs).

On *T* trials, values of summed similarity arise from two quantitatively different sources that differ in their respective variability. The first, far larger source of variability reflects the contribution of the *n*−1 study items that are not replicated by **p**. Random selection of study items from a stimulus pool means that some of *n*−1 study items will be similar to **p**, and that others will be very different from **p**. As a result of this random divergence, these *n*−1 nonmatching study items will contribute a highly variable amount of similarity to the summed-similarity signal for any trial. The second, smaller source of variability in summed similarity on *T* trials reflects the contribution of the one study item that the **p** does replicate. Over trials, this study item's representation will tend to be perceptually similar to **p**—even with the memorial noise postulated by the model. Because that study item and **p** are physically identical, they are likely to be perceptually similar, despite the random noise associated with the study item's memorial representation. As a result, similarity between this study item and **p** will vary over a narrow range clustered near 1.0 (Zhou, Kahana, & Sekuler, 2004).

R. Sekuler, sekuler@brandeis.edu

On L trials, variability in the summed-similarity signal arises from n study items that do not replicate $\mathbf{p}$ (by definition, there is no item in the study list that replicates $\mathbf{p}$). Again, random selection of study items from some pool means that some of the n study items will be very similar to $\mathbf{p}$, and others will be quite different. As a result, on L trials, the n study items will each make a highly variable contribution to each trial's summed-similarity signal. In contrast, on T trials, $n-1$ study items make a highly variable contribution to summed similarity, whereas the remaining study item contributes little variability. Adding the sources of variability for the two trial types, the variance in the summed-similarity signal will be greater for T trials than for L trials.

If subjects' recognition judgments were based on summed-similarity values, then the ratios of expected variances on T and on L trials would produce a ROC with a telltale characteristic. This characteristic would be most easily seen when $P(\text{yes}|T)$ and $P(\text{yes}|L)$—the ROC's x and y values—were transformed to standard z scores and then replotted. In this new plot, the slope of the transformed ROC—known as a zROC—would reflect the ratio of the variances of the summed similarities on T trials (in the denominator) and L trials (in the numerator). In order to test this prediction that zROC slopes will be greater than 1.0, subjects in two experiments expressed their judgments using an analogue rating scale. Converting the analogue judgments into rating-scale equivalents facilitated the generation of zROCs, whose slopes could be compared with predicted values. We should note that the predicted zROC slopes greater than 1.0 would not be consistent with prior results from studies with verbal materials presented in lists much longer than ours (see, e.g., Donaldson & Murdock, 1968; Murdock, 1982). Those studies produced zROCs with slopes less than 1—a finding that has resisted reconciliation with existing theory (Ratcliff, Sheu, & Gronlund, 1992).

In order to test these predictions, we examined short-term memory, using both old–new recognition judgments (asking subjects to judge whether some probe item [$\mathbf{p}$] had been part of a just-presented list of three study items) and item-position identifications (asking subjects to judge the serial position of the study item that matched $\mathbf{p}$). On each trial, subjects saw three briefly presented study items. The series of study stimuli—whose members varied from trial to trial—was followed by a probe item ($\mathbf{p}$), which either replicated one of the preceding study items or differed from all three. Subjects used an analogue rating scale to identify the serial position whose study stimulus matched $\mathbf{p}$; if no study item matched $\mathbf{p}$, then subjects registered that judgment with a *no* response. The analogue scale also allowed subjects to express their confidence that they had made a correct response (Watson, Rilling, & Bourbon, 1964). These expressions of confidence were used in generating ROCs, which served as one of our analytic tools.

EXPERIMENT 1

In Experiment 1, 75% of all trials were target (T) trials, with the probe item replicating one of the three study items. Additionally, we use t_i to designate the serial posi-

tion in which the study item was replicated by $\mathbf{p}$. On T trials, a random number generator assigned t_i with equal probability to the first, second, or third serial position in the set of study items. Allowing T trials to occur three times as often as L trials facilitated examination of possible effects of t_i's serial position.

Method

Stimuli. Stimuli for each trial were drawn from a pool of compound sinusoidal gratings; each comprised superimposed vertical and horizontal sinusoidal luminance gratings. Each component's contrast was set to 0.2—a value well above detection threshold. Edge effects were minimized by windowing the grating with a circular 2-D Gaussian (space-constant 1° of visual angle). Before the window was applied, a grating's width was 5° of visual angle.

In order to reduce the influence of individual differences in perception on measurements of memory performance, each subject's memory stimuli were scaled according to that subject's visual discrimination threshold (Zhou et al., 2004). Doing so produced for each subject a unique pool of compound grating stimuli with five vertical spatial frequencies crossed with five horizontal spatial frequencies. Frequency discrimination thresholds ranged from 4.3% to 13.1% ($M = 8.2$, $SD = 3.1$). Vertical as well as horizontal spatial frequencies were 2 cpd, plus or minus three or six times a subject's Weber fraction for spatial frequency. Since the mean Weber fraction was 0.082, the spatial frequencies of the average stimuli were 1.51, 1.75, 2.0, 2.25, and 2.49 cpd. Figure 1 illustrates the set of compound gratings that correspond to these typical values. Enforcing a minimum between-stimulus difference of three times a subject's discrimination threshold reduced the likelihood that perceptual confusions between pairs of stimuli with minimal memory load would by themselves lead to misidentifications. A supplementary experiment with five additional subjects showed that when just two gratings were presented for a same–different comparison and those gratings' spatial frequencies differed by three times the subject's discrimination threshold, perceptual confusions were rare—occurring on fewer than 3%–4% of the trials.

Subjects. Subjects were 10 paid volunteers whose ages ranged from 19 to 28 years ($M = 22.9$, $SD = 3.3$). Subjects' acuity (measured with Landolt C targets) ranged from 20/13 to 20/22 ($M = 20/16.8$, $SD = 2.7$); contrast sensitivity (measured with the Pelli–Robson charts; Pelli, Robson, & Wilkins, 1988) ranged from 1.80 to 1.95 ($M = 1.92$, $SD = 0.06$).

Procedure. Each trial's set of study items comprised three compound gratings, followed by a probe stimulus ($\mathbf{p}$). Each of the three study stimuli ($\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$) was presented for 750 msec, separated by interstimulus intervals of 400 msec each. Then, after a delay of 1,000 msec, a warning tone sounded, and $\mathbf{p}$ was presented for 750 msec. One second later, a response scale was presented, and it remained visible until the subject's response had been registered. Subjects used the response scale shown in Figure 2A to report whether $\mathbf{p}$ had been in the study set and, if so, which study item (first, second, or third) was matched by $\mathbf{p}$. This scale consisted of four selection arms that were labeled *None*, *First*, *Second*, or *Third*. If $\mathbf{p}$ seemed to match one of the study items, then subjects used the computer mouse to identify the arm that corresponded to the serial position of the study stimulus ($\mathbf{s}_1$, $\mathbf{s}_2$, or $\mathbf{s}_3$) that matched $\mathbf{p}$. If $\mathbf{p}$ seemed to match none of the study items, then the subject positioned the cursor on the arm labeled *None*. In addition, subjects were encouraged to position the cursor in a way that expressed their confidence that they had selected the response-correct arm. In particular, cursor positions near the intersection of the four arms signaled little confidence in the judgment; positions near an arm's outer end signaled high confidence. Since subjects' responses reflected their confidence that they had identified the correct serial position, the task qualified as a Type I rating task (see, e.g., Macmillan & Creelman, 2005).

Once the subject was satisfied with the cursor's location within the response arm, a click of the computer mouse button caused the

Horizontal Spatial Frequency (cycles/degree)
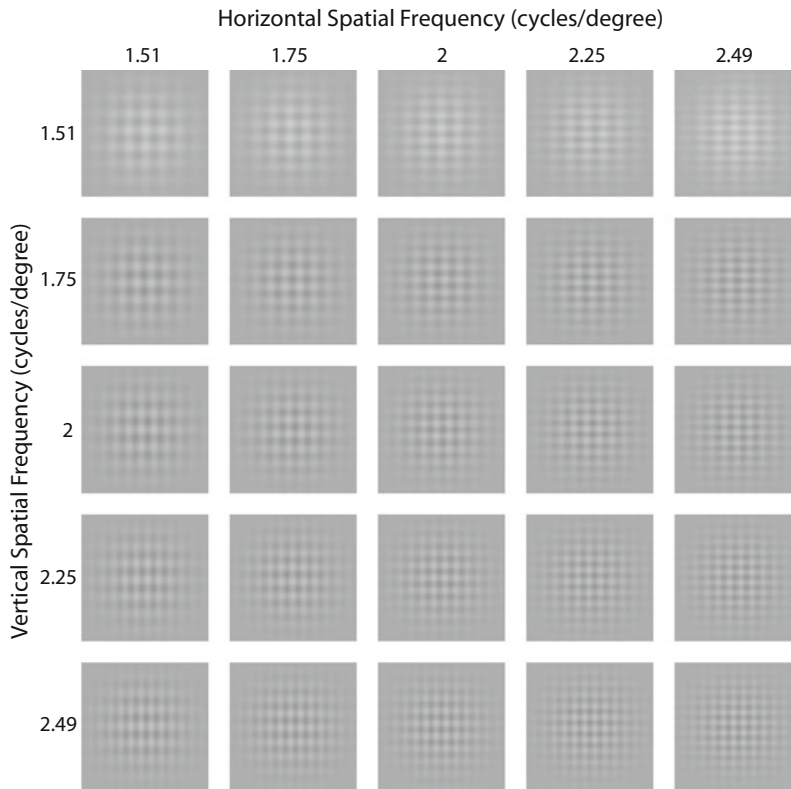
Figure 1. The average set of stimuli used in each experiment. Within each row, vertical spatial frequency changes by three or six threshold units, decreasing and increasing from the mean of 2.0 cpd, shown in the center of the stimulus matrix. Within each column, horizontal spatial frequency changes in the same way, again relative to the mean of 2.0 cpd.

computer to register the cursor's location. No instructions were given about the speed with which subjects should respond. On average, once the scale was presented, a response was registered in about 2–3 sec, which was sufficiently short that memory would not have decayed significantly (Kahana & Sekuler, 2002; Sekuler, Kahana, McLaughlin, Golomb, & Wingfield, 2005).

Distinctive tones provided feedback about response correctness. On *T* trials, feedback was contingent on the response's identification component: Feedback signaled whether the subject's response cor-

rectly identified which study item (s_1, s_2, or s_3) matched **p**; as with an incorrect identification response, a *None* response on a *T* trial brought feedback that the response was wrong. On *L* trials, feedback was contingent on whether the response correctly reflected that none of the study items matched **p**; all other responses—*First*, *Second*, or *Third*—were followed by feedback that the response had been wrong.

The display's mean luminance was maintained at 17.8 cd/m², which prevented distracting luminance transients that would otherwise have accompanied a change of stimulus. A subject viewed the
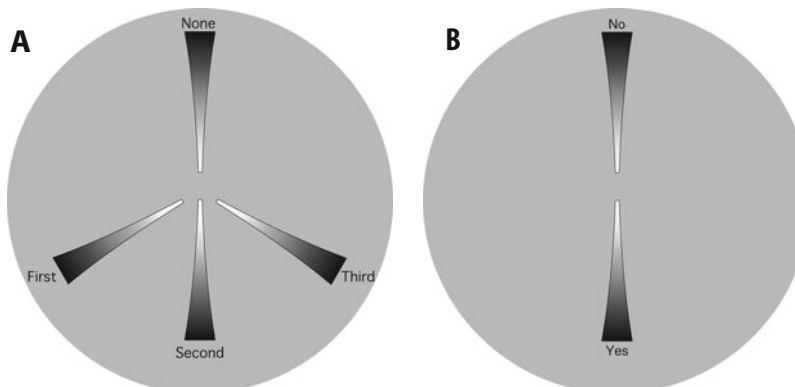
Figure 2. (A) The four-armed analogue selection scale used by subjects to express judgments in Experiments 1 and 2. (B) The two-armed analogue selection scale used to express recognition judgments in Experiment 3.

stimulus display from a distance of 114 cm, the head supported and steadied by a combination head rest and chin cup. Trials were self-paced. On each trial, s_1, s_2, and s_3 were sampled randomly without replacement from the pool of 25 stimuli that had been generated for that subject. On 75% of the trials, p replicated s_1, s_2, or s_3 with equal frequency. On the remaining trials, p was chosen randomly from the 22 members of the stimulus pool that were not among that trial's three study items. These probabilities were explained to subjects prior to the experiment. Each subject was tested on 800 trials that were distributed across four 1-h sessions.

EXPERIMENT 2

Although Experiment 1 examined item-position identification, its overall design was grounded in prior studies of recognition memory in which no identification responses were taken (Kahana & Sekuler, 2002; Kahana et al., 2007). Because those studies used balanced schedules of T and L trials, we were concerned that Experiment 1's unbalanced schedule might undermine comparisons between Experiment 1 and previous studies. Therefore, Experiment 2 replicated the conditions of Experiment 1, but with a balanced schedule of T and L trials.

Method

Subjects. Five paid volunteers (age 18–21 years, $M = 19.8$, $SD = 1.1$) participated in this study. None had served in the preceding experiment. Subjects' acuity ranged from 20/13 to 20/20 ($M = 20/16.6$, $SD = 2.7$); contrast sensitivity ranged from 1.80 to 1.95 ($M = 1.89$, $SD = 0.08$); frequency discrimination thresholds ranged from 6.0% to 10.2% ($M = 9.2$, $SD = 3.1$). All measurements used the same techniques as those employed in the previous experiment.

Stimuli and Procedure. This experiment used the same family of stimuli as those in the previous experiment, with stimulus spatial frequencies again tailored to individual subjects' frequency discrimination thresholds. The proportion of T trials was reduced from 75% in Experiment 1 to 50% in Experiment 2. As before, with equal probability, p matched s_1, s_2, or s_3. These probabilities were explained to the subjects prior to the experiment. There were no other differences between this experiment and its predecessor. Each subject was tested on 800 trials, distributed across four 1-h sessions.

EXPERIMENT 3

In the preceding two experiments, subjects made an item-position judgment on each trial, either identifying the serial position of the study item that matched p, or responding *None*. It was computationally simple to transform those item-position judgments into equivalent recognition memory, but we cannot ignore the possibility that the result might not truly correspond to recognition measured directly. How well does recognition that is measured indirectly, by means of transformed item-position judgments, correspond to recognition measured when no item-position judgment was required? In order to answer this question, for this experiment, we modified the task used in the preceding two experiments, eliminating position judgments and requiring only recognition judgments.

Method

Subjects. Five paid volunteers (age 18–20 years, $M = 18.6$, $SD = 0.9$) participated. None had served in the preceding experiments. Subjects' acuity ranged from 20/15 to 20/25 ($M = 20/18$, $SD = 4.5$); mean contrast sensitivity was 1.95; frequency discrimination thresh-

olds ranged from 6.6% to 18.7% ($M = 10.6$, $SD = 4.7$). Measurements were made using the same techniques as those used in the previous experiments.

Stimuli and Procedure. The only difference between Experiment 2 and Experiment 3 was the judgment required of subjects and the response selection screen on which judgments were registered. As shown in Figure 2B, for Experiment 3, the oblique arms of the four-arm response display were eliminated, and subjects indicated only whether p had or had not been among the study items for that trial; no identification response was required. To remind subjects of the task, one arm of the response screen was labeled *yes*, and the other arm was labeled *no*. As before, participants signaled their confidence in each judgment by clicking on the appropriate arm, with distance from the center of the screen indicating increasing confidence. Subjects were informed that T and L trials would occur with equal frequency. In other respects, Experiment 3 was identical to Experiment 2. Each subject was tested on 800 trials that were distributed across four 1-h sessions, and stimuli were tailored to individual subjects' frequency discrimination thresholds.

RESULTS

Overall performance

Experiment 1. Figure 3A shows the proportion of correct old–new recognition and item-position identification in Experiment 1. The proportion correct for L trials ($.51 \pm .03$) is shown by the single data point at the panel's right side. The three serial position curves—one for each response measure—show a strong recency effect, with performance on T trials improving from s_1 through s_3. A repeated measures ANOVA confirmed this result, showing a significant main effect of serial position for both curves in Figure 3A [$F(2,18) = 26.17$, $p < .01$]. False alarm rates—the proportion of identifying L as T—were .12, .13, and .11 for s_1, s_2, and s_3, respectively. These false alarm rates did not differ reliably across serial positions [$F(2,18) = 1.60$, $p > .20$].

Experiment 2. Figure 3B shows that the proportion of correct responses for L trials was $.55 \pm .03$. Additionally, as in Experiment 1, the serial position of the study item matched by p had a substantial effect [$F(2,8) = 20.70$, $p < .01$]. False alarm rates were .09, .15, and .09 for s_1, s_2, and s_3, respectively.

Experiment 3. Recognition memory performance for Experiment 3 was expressed as the proportion of correct recognition responses ($.70 \pm .26$). As with the previous experiments, a repeated measures ANOVA demonstrated a main effect for the serial position of the study item that matched p [$F(2,8) = 11.62$, $p < .01$]. We can compare the correct recognition measures from Experiment 2—in which recognition was calculated by summing identification responses—to the recognition measure from this experiment. A repeated measures ANOVA showed that the serial position results for Experiments 2 and 3 did not differ from one another [$F(2,16) = 1.74$, $p > .20$].

ROC Analysis

In order to compare recognition as measured indirectly (by identification judgments; Experiments 1 and 2) and recognition measured directly (Experiment 3), we calculated the proportion of correct recognition responses produced in the two tasks. Recognition measures differed
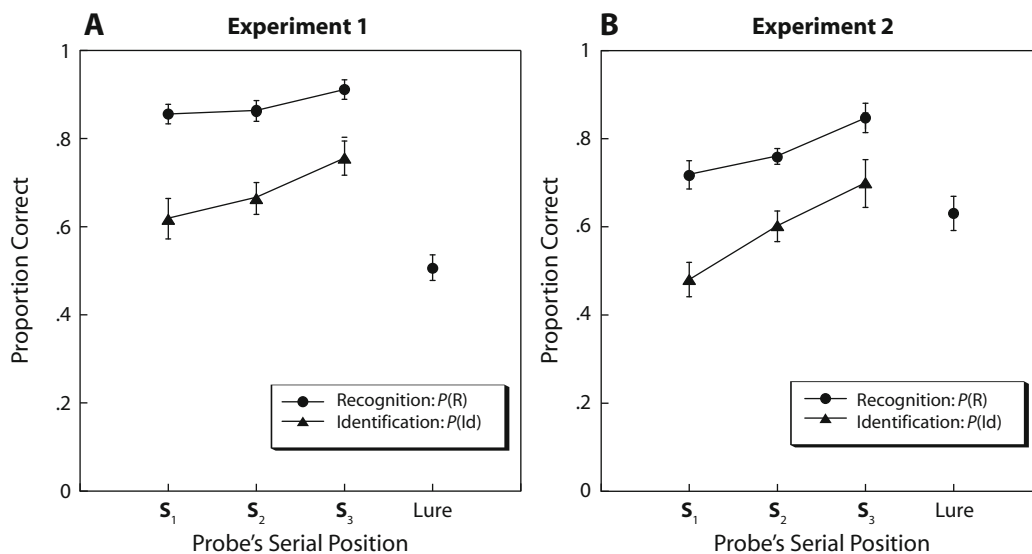
Figure 3. Proportion correct recognitions and identifications as a function of the serial position of the study item replicated by p. Also shown is the proportion correct rejection of lure stimuli (right side of each panel). Vertical bars around each data point represent ±1 within-subjects standard error (see, e.g., Loftus & Masson, 1994).

significantly between Experiments 1 and 2 ($p < .02$), most likely because of the experiments' different ratios of T to L trials. Given that T trials comprised 75% of all trials in Experiment 1 but just 50% of the trials in Experiments 2 and 3, signal detection theory would predict these two values of $P(R)$ to be ordered as they are. This hypothesis is bolstered by the observation that the overall proportion of both correct recognitions and false alarms was higher in Experiment 1 than in Experiment 2 (see Figure 3).

The difference in stimulus schedule could have affected performance by changing the accuracy of memory (such as what might come from differences in task difficulty and attentional demands), from a change in subjects' criteria, or from some combination of the two. In order to choose among these alternatives, we generated ROC curves from the judgments in each of the three experiments. In doing this, we exploited the confidence judgments provided by subjects' use of the continuous, analogue rating scale (Nachmias & Steinman, 1963).

Because our rating scale was continuous rather than categorical, we sought an empirical estimate of how many useful categories—variations in confidence—were actually represented in subjects' use of the analogue scale. After estimating that number, we used it to set the number of categories used to generate ROC curves from subjects' expressions of confidence. In order to determine the useable grain of the analogue rating scale, we partitioned the rating scales into varying numbers of bins and, for each number, we calculated the amount of information transmitted by the responses (Garner & Hake, 1951). Watson et al.'s (1964) method was used to calculate information transmitted for correct identification responses that had been sorted post hoc into varying numbers of bins. The number of post hoc response bins was varied from 2 to 12, with the constraint that for any subject, all bins contained equal numbers of responses. The information transmitted by these correct responses grew with the number of response categories, but approached asymptote with no more than 10 response categories. Therefore, in generating ROC curves, we partitioned the analogue confidence responses into 10 categories, with equal numbers of responses in each (see Nachmias & Steinman, 1963). In order to test the similarity of results in Experiments 1 and 2, individual ROC curves were generated for each subject, and the area under each curve was calculated using the trapezoidal rule for numerical integration (Wickens, 2002). Because we anticipated that our data would violate the equal-variance assumption, areas were calculated as values of A_z, rather than of A' (Wickens, 2002). In these calculations, the few response bins whose cumulative proportions were either 0 or 1.0 were dropped.

Figure 4A shows mean ROC curves for the three experiments. In generating ROC curves for Experiments 1 and 2, the analogue confidence ratings for r_1, r_2, and r_3 were aggregated. For the ROC curve derived from Experiment 3, we used the analogue confidence ratings that were associated with *yes* and *no* responses. ROCs were generated for each subject, and the area under each subject's ROC curve was computed. The mean area, A_z, under the ROC and the standard errors associated with that mean was 0.73 ± 0.019, 0.75 ± 0.027, and 0.68 ± 0.037, for Experiments 1, 2, and 3, respectively. We cannot explain why performance in Experiment 3 was somewhat worse than that in the other experiments, and the small number of subjects in Experiments 2 and 3 make us hesitant to speculate about this point.

A one-way ANOVA confirmed that areas under the ROCs for the three experiments did not differ significantly from one another [$F(2,19) = 1.56$, $p > .20$]. This outcome suggests that recognition as measured directly in Experiment 3 is well approximated by recognition as estimated by aggregating over the three separate identification responses: r_1, r_2, and r_3.
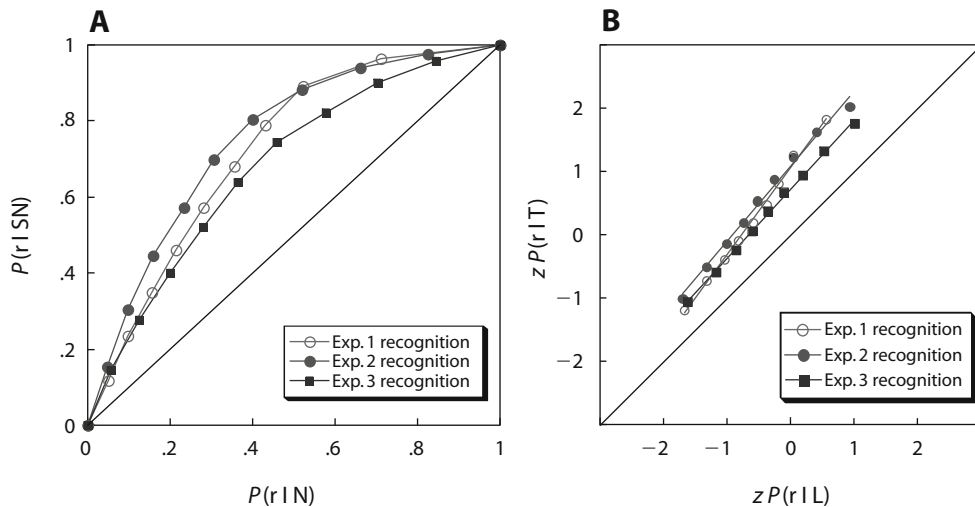
Figure 4. (A) Receiver operating characteristics (ROCs) for recognition performance in Experiments 1–3. ROCs for Experiments 1, 2, and 3 are represented by open circles, filled circles, and filled squares, respectively. (B) zROC curves generated from recognition performance in Experiments 1, 2, and 3 are represented by open circles, filled circles, and filled squares, respectively. A zROC is a ROC in z coordinates, where z is a transformation that converts a proportion into its corresponding z or standard score.

Item-Position Errors

In both Experiments 1 and 2, subjects made many misidentifications of serial position. On about 25% of all T trials, subjects correctly rejected the *no* response, only to misidentify the serial position of the study item that had actually been replicated by $\mathbf{p}$.

In order to evaluate possible causes of misidentifications, we asked whether misidentifications might be explained by some structural or positional attribute of the stimuli. The structural attribute was the pairwise spatial similarity between exemplars in the two-dimensional Euclidean space within which our stimuli were defined. This variable's likely potency is suggested by a summed-similarity account of false alarms and recognition judgments. For the positional attribute, imagine that there was some orderly, nonrandom forgetting of serial position information. One form of nonrandom forgetting produces a "locality constraint," which promotes local rather than global errors. In the task at hand, items that occupied serially adjacent positions in a study sequence would be more likely to be confused with one another than would positions that were more widely separated in a sequence (Lee & Estes, 1977; Page & Norris, 1998). More specifically, with partial loss of serial position information, $\mathbf{s}_1$ would be more likely to be misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_3$, and $\mathbf{s}_3$ would be more likely to be misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_1$. This account is mute on errors arising from forgetting $\mathbf{s}_2$'s serial position in a three-item list like those used here. Note that the duration of each study item (750 msec)—together with the 400 msec separating successive items—should have been sufficient to minimize perceptual confusions between intervals, which suggests that misidentifications arise from failure of memory rather than failures of perception.

In order to evaluate competing accounts of misidentifications, each participant's item-position errors were sorted into the cells of a 2×2 matrix. In populating the cells of the matrix, we considered only trials on which $\mathbf{p}$ actually matched either the first or last study item—$\mathbf{s}_1$ or $\mathbf{s}_3$; the remaining trials, on which $\mathbf{p}$ matched $\mathbf{s}_2$, did not lead to unequivocal predictions for misidentifications. The matrix's rows corresponded to two levels of a variable we call *structural similarity*; the matrix's columns correspond to two levels of a variable we call *positional similarity*. In order to generate the value of spatial similarity, we used a metric stimulus space in Figure 1 to calculate the Euclidean distance in spatial frequency between (1) $\mathbf{p}$ and the misidentified study item, and (2) $\mathbf{p}$ and the remaining study item that did not match $\mathbf{p}$. If the first of these two distances was smaller, then we categorized structural similarity between $\mathbf{p}$ and the misidentified item as "high"; otherwise, we categorized structural similarity as "low." For positional similarity, we categorized misidentifications according to whether the error in identification represented a shift of either one (high similarity) or two (low similarity) serial positions. For example, if $\mathbf{s}_2$ were misidentified as matching $\mathbf{p}$ when the actual matching study item was $\mathbf{s}_3$, then this error of one serial position was categorized as high positional similarity; if $\mathbf{s}_1$ were misidentified as matching $\mathbf{p}$ when the actual matching study item was $\mathbf{s}_3$, then the error of two serial positions was categorized as low positional similarity. A factorial cross of structural and positional variables produced four combinations of differences between $\mathbf{p}$ and the misidentified study item. Trials involving a match to $\mathbf{s}_2$ were omitted from this analysis, because univocal predictions for those trials could not be made within our theoretical framework.

Figure 5 shows the proportions of item-position errors in Experiments 1 and 2 that fell into each of the four categories. In each panel, the horizontal axis shows the two levels of spatial similarity; black bars show results with

high positional similarity, and gray bars show results with low positional similarity. Note first that we can easily dismiss the hypothesis that all misidentifications resulted from a completely stochastic process. Monte Carlo simulation showed that a completely random process, which would produce correct identification of serial position on just .1875 of all trials, would generate values as extreme as the highest value in either panel of Figure 5 on fewer than 1 in 100,000 replications of the experiment. In fact, the distribution of misidentifications is consistent with the alternative hypothesis—namely, that both positional and physical similarity induced identification errors, with physical similarity producing a larger effect than did positional similarity. Moreover, the two panels in Figure 5 show no evidence of an interaction between the two variables; that is, the black and gray bars at low structural similarity differ by about as much as the corresponding bars at high structural similarity in both Experiments 1 and 2. Finally, there is some hint in Figure 5 that positional similarity might have played a larger role in errors made in Experiment 2 than in Experiment 1. Although it is tempting to compare effect sizes across experiments, the relatively small samples of subjects make us hesitant to attempt statistical comparisons of this kind.

The distribution of misidentifications across the four categories in the notional 2×2 matrix suggests that both positional and structural similarity induced position errors, with structural similarity exerting a larger effect than did positional similarity (compare the pair of bars at the left side of Figure 5A with those corresponding at the figure's right). The figure shows no evidence of an interaction between the two variables; that is, the black and gray

bars at the left side of Figure 5A differ by about as much as the corresponding bars at the panel's right side.

Potentially, some errors could have arisen from purely perceptual confusions among stimuli, with no actual involvement of memory per se. However, stimulus series were constructed so that any two stimuli differed in spatial frequency by at least three times the subject's difference threshold. As was noted earlier, when stimuli like ours differ by that much, perceptual confusion alone with minimal contribution from errors in memory would have caused stimuli to be mistaken for one another only 3%–4% of the time. Thus, this source of mistaken identity does not account for the much higher proportion of misidentifications actually obtained in the memory experiment.

Previous studies showed that summed similarity of $\mathbf{p}$ and study items is an effective predictor of recognition memory (Kahana & Sekuler, 2002; Nosofsky, 1992). Earlier in this article, we described how a summed-similarity computation can explain the origin of some misidentifications. As it is explained below, summed-similarity computations also make a novel prediction for a key property of the ROCs.

Figure 7 illustrates in schematic form some key elements of NEMo, a summed-similarity model for recognition proposed by Kahana and Sekuler (2002). Although our exposition here revolves around NEMo, we should note that many of the same points could be made using other global matching models, which share similar computations (see, e.g., Brown, Neath, & Chater, 2007; Estes, 1986; Lacroix, Murre, Postma, & van den Herik, 2006; Lamberts, Brockdorff, & Heit, 2003; Shiffrin & Steyvers, 1997). NEMo assumes that study items $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$ are stored in memory
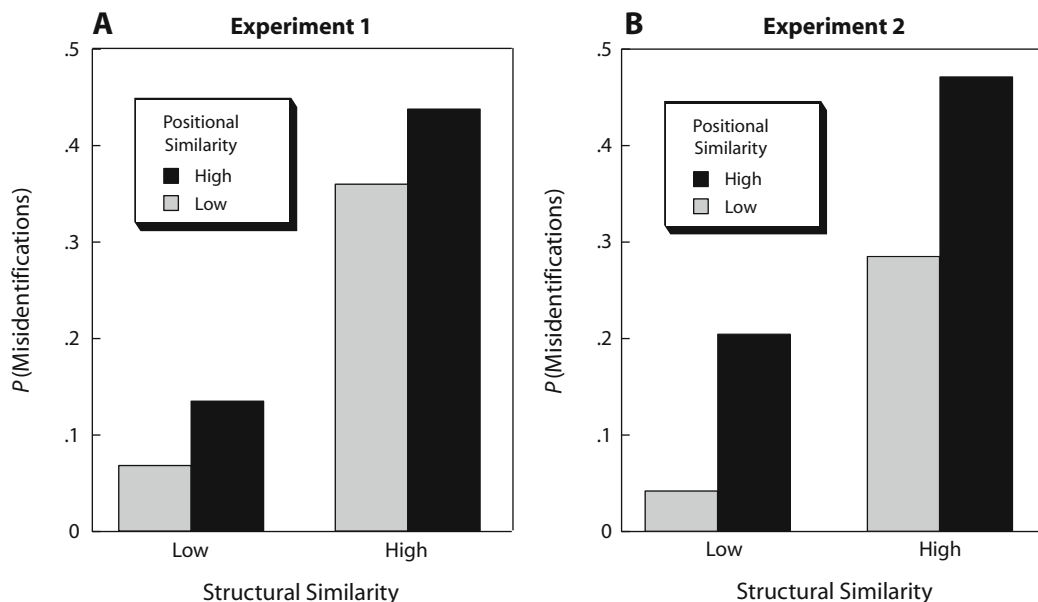
Figure 5. Proportion of misidentified items on *T* trials as a function of the structural and positional similarity between correct and misidentified study stimuli. (A) Misidentifications in Experiment 1. (B) Misidentifications in Experiment 2. Two levels of structural difference are plotted on the *x*-axis. Each of them is plotted separately for stimulus pairs that were similar in position (black bars) and for those that were dissimilar in position (gray bars).
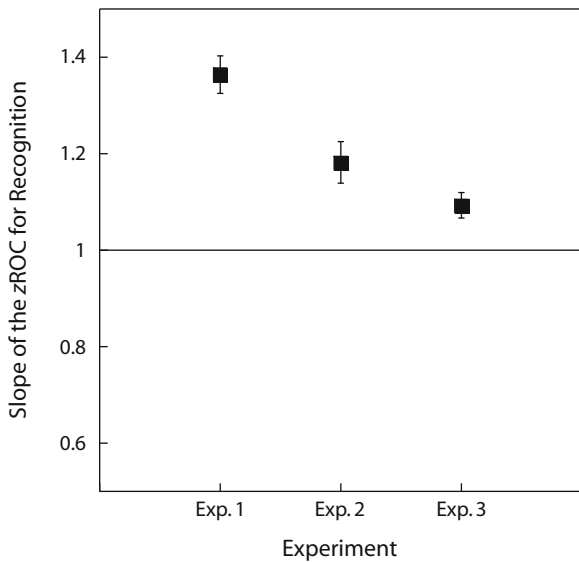
Figure 6. Mean slopes of *z*ROCs measured in each experiment. An error bar represents ±1 within-subjects standard error (see Loftus & Masson, 1994).
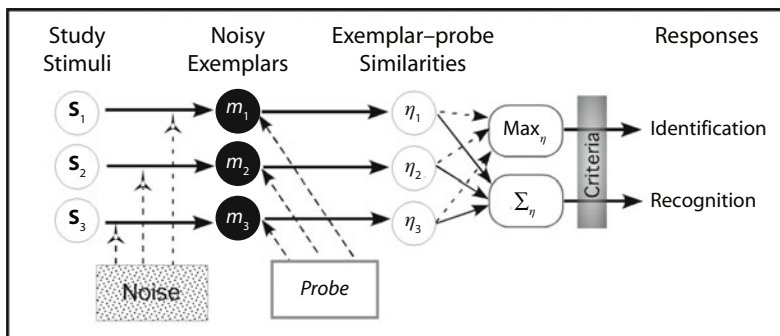
as corresponding noisy exemplars—$\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$—where exemplars' subscripts signify the order in which the stimuli were presented. NEMo computes pairwise similarities—η_1, η_2, η_3—between $\mathbf{p}$ and the noisy exemplar of each study item. If the sum—$\Sigma\eta$—of these pairwise similarities exceeds an (optimal) criterion, then the model responds that $\mathbf{p}$ had been in the study series.[1] If $\Sigma\eta$ fails to exceed the criterion value, then the model responds that $\mathbf{p}$ was not among the items in the study series. In NEMo's computation, sets of study and $\mathbf{p}$ items—together with random noise in the exemplar representations—produce a distribution of values of $\Sigma\eta$. By definition, on T trials, $\mathbf{p}$ must

be a physical replica of one study item; on L trials, none of the study items is replicated by $\mathbf{p}$. Thus, even in the presence of exemplar noise that may reduce the remembered similarity of $\mathbf{p}$ and the study item it replicates, $\Sigma\eta$ will tend to be larger on T trials than on L trials, which will cause the proportion of recognition responses ("hits") to differ on the two trial types. False recognitions ("false alarms") may occur either because on some L trials, exemplar noise will cause $\Sigma\eta$ to exceed the subject's criterion, or because the lure is similar to many of the study-list items.

Predicting the Form of the ROC

In order to make predictions about the form of ROC curves, we cast the distributions of $\Sigma\eta$ on T and on L trials into signal detection terms, treating the distribution of $\Sigma\eta$ for T trials as the signal distribution and the distribution of $\Sigma\eta$ for L trials as the noise distribution. Within a signal detection framework, then the slope of a linear zROC curve reflects the relative variances of the $\Sigma\eta$ distributions on T and on L trials (Wickens, 2002). If the two distributions' variances were equal, then the resulting zROC's slope would be 1. But, as was explained earlier, a model-based account predicts that variance in $\Sigma\eta$ will be larger on L trials than on T trials. Hence, the zROC slopes should be greater than 1 (Wickens, 2002). In order to test this prediction, zROC curves were generated for each subject by cumulating hit and false-alarm rates over response bins (as was described earlier) and converting the cumulated values into standard scores. The mean zROC curves for each experiment are shown in Figure 4B.

For each experiment, the mean zROC curve generated by averaging zROCs from individual subjects is well described by a linear function. Values of r^2s for the linear terms in a second order polynomial fit were $>.95$; the addition of a quadratic term improved the fit by less than .02, which was not statistically reliable. In a signal detection

Figure 7. Schematic of model showing key stages that could lead to recognition and identification responses. Samples of noise are added to visual representations of the study stimuli (s_1, s_2, s_3) producing a set of corresponding, noisy exemplars (m_1, m_2, m_3) that are stored in memory. At the presentation of the probe stimulus, the similarity between p and each memory representation is computed and stored as η_1, η_2, and η_3. These separate similarity measurements are combined into a summed similarity value, $\Sigma\eta$. Omitted from this schematic representation are parameters that transform physical stimulus distances into perceptual similarity, and β, which captures variation in the overall similarity of study items to one another. So that the model can also identify the serial position of the study item that matched p, one might add a max operator that returns the serial position associated with the largest member of the similarity set, η_1, η_2, η_3.

framework, the linearity of zROC curves is consistent with underlying distributions that are normal (Murdock, 1982). The mean zROC slopes for each experiment are shown in Figure 6. The zROCs were generated from recognition performance measured directly (as in Experiment 3) or indirectly (for Experiments 1 and 2). Each had a slope significantly greater than 1; this was true even for Experiment 3, whose zROCs had the lowest slopes [$t(4) = 3.45$, $p < .03$]. The slopes from Experiments 2 and 3 were not significantly different from one another ($p = .40$). Finally, as Table 1 shows, the obtained slopes did not vary systematically with serial position when zROCs were computed for each serial position separately. This consistency of the zROC slopes across the three serial positions suggests that subjects were probably not using substantially different strategies to recognize matching items that occupied different serial positions (Malmberg & Xu, 2006).

In order to verify the link between zROC slope and the distributions of $\Sigma\eta$ values on T and L trials, we calculated the summed similarity for each trial in our study. In doing so, we substituted pairwise Euclidean differences in spatial frequency for the corresponding perceptual differences that are usually used in NEMo. Although these two variables—physical and perceptual distance—are likely related by an exponential transform (Shepard, 1987; Yotsumoto, Kahana, Wilson, & Sekuler, 2007), any increasing monotonic relationship between the two variables would leave the argument unchanged. Summed **p**–study distances were calculated separately for both T trials and L trials in Experiment 3.

The frequency distributions of $\Sigma\eta$ for all stimulus sets that appeared in Experiment 3 are plotted in Figure 8A. Note that the x-axis has been reversed so that the smallest value of summed distance lies to the right. Because summed similarity—$\Sigma\eta$—and summed distance are inversely related, the reversal of the normal x-axis direction represents increased summed similarity from left to right and also brings the visual format of Figure 8A's distributions into conformity with formats commonly used in signal detection theory. As was expected, the mean value of $\Sigma\eta$ for T trials tended to be larger than the comparable value for L trials; also, the distribution of $\Sigma\eta$ for L trials had larger variance than did $\Sigma\eta$ for T trials. This unequal variance in the empirical signal and noise distributions predicts zROCs with slopes greater than 1 (Wickens, 2002).

In order to confirm that distributional differences in $\Sigma\eta$ for T and L trials would actually have the predicted effect on zROC slope, we used a signal detection approach to simulate the recognition data that would be produced by several different distributions of $\Sigma\eta$. In particular, we simulated recognition for several different subsets of stimuli,

all of which were derived from the full set of stimuli used in Experiment 3—the one experiment in which recognition was measured directly. First, we calculated the value of $\Sigma\eta$ for all 400 L trials and for all 400 T trials that each subject saw. Figure 8A's lower panel shows the distributions of $\Sigma\eta$ for the two trial types, aggregated over all subjects. Note that because stimulus sets were generated randomly from the pool of 25 items, each subject had been tested with a partially unique set of T and L stimuli. From each subject's responses to the 400 T and 400 L trials, we generated a zROC and calculated its slope. Note that these 800 trials were the actual stimulus set used in Experiment 3; the mean slope shown in the leftmost point in Figure 8B was 1.09—a value identical to the corresponding value obtained in the actual experiment (see rightmost point in Figure 6B). Note that not all characteristics of the distributions shown in Figure 8 are faithful to what NEMo assumes. In particular, for the sake of transparency, we have not applied a similarity transformation to the physical stimuli, choosing instead to work with Euclidean distances between the spatial frequencies of our stimuli.

Having confirmed that the variances of distributions of $\Sigma\eta$ differed between the complete set of our T and L trials, we constituted two subsets of stimuli. For the first subset, we sought to reverse the relationship between the distributions' variances while holding constant the difference between the distributions' means; we reasoned that these new modified distributions should produce zROC slopes less than 1. Starting with each subject's original stimulus set of 400 T and 400 L trials, we carved out a reduced-variance L distribution by selecting 60 L trials whose $\Sigma\eta$ values were near the mean value of $\Sigma\eta$. In addition, 60 T trials were randomly selected without regard to their value of $\Sigma\eta$. This maneuver reduced the variance for noise trials, leaving the variance for signal trials unchanged while also preserving the mean difference between the two distributions. Altering the relative variances of $\Sigma\eta$ for T and L trials had the expected effect on zROC slope: The relatively narrower distribution of summed similarities values on L trials produced a simulated zROC mean slope of 0.85—a value considerably below 1.0 (the rightmost value shown in Figure 8B).

Finally, we did simulations with a somewhat larger subset of values sampled from the original distribution of $\Sigma\eta$ from L trials; the aim was to generate a sample of stimuli in which the variance of $\Sigma\eta$ for L trials was approximately the same as that for T trials, which should produce zROC slopes very near 1.0. We drew 250 L trials and 250 T trials from the original sets of 400 items. L trials were drawn without replacement from a region in the vicinity of the mean for all L trials, but T trials for the subset were drawn at random from the entire distribution of 400 trials. The result was a ratio of variances between distributions that was intermediate to the ratios for the 60-trial sets and for the complete 400-trial sets. Again, in constructing this set of stimuli, we held constant the mean distance between these two distributions of $\Sigma\eta$. From each subject's responses in Experiment 3 to each of these stimulus lists, we generated a zROC for that subject. The mean and standard error of the zROC slopes are shown by the middle point in Figure 8B. Note that as expected, the mean zROC slope here

Table 1
Mean zROC Slope and Standard Error
for Each Serial Position and Experiment

	Position 1		Position 2		Position 3	
	M	SE	M	SE	M	SE
Experiment 1	1.26	0.02	1.29	0.02	1.32	0.02
Experiment 2	1.13	0.01	1.12	0.07	1.19	0.02
Experiment 3	1.08	0.03	1.04	0.07	1.08	0.03
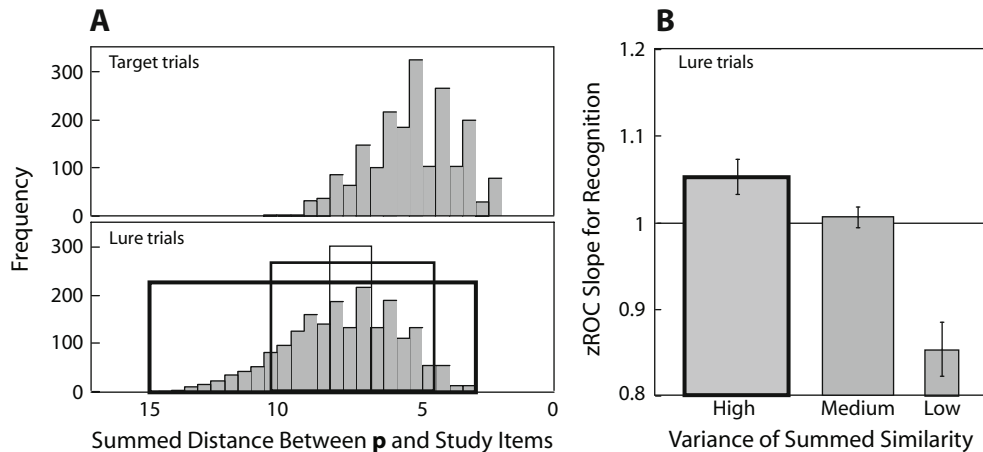
Figure 8. (A) Distributions of summed similarity values for all *T* trials (upper subpanel) and all *L* trials (lower subpanel) used in Experiment 3 and in simulations of recognition. *L* trials used in both Experiment 3 and the first simulation are bracketed by the rectangle with the thickest line. The two narrower rectangles bracket *L* trials that were used in the second and third simulations. **(B)** Mean and *SEM* slopes of *z*ROCs calculated for three sets of *L* trials; the variance of *L* trials' summed similarity values varied systematically among the three sets of trials. The rightmost point shows the mean and *SEM* *z*ROC slope simulated for all stimulus sets used in Experiment 3. The thick, medium, and thin bars each correspond to the simulations with large, medium, and small variance of summed similarity. The thickness of lines in panel B corresponds to that of those in panel A. The standard deviation of the *T* distribution is 1.68; the standard deviation for *L* trials was 2.1.

was intermediate to those from the other two simulated conditions and was close to 1.0.

The simulations represented in Figures 8A and 8B confirm that the slopes of *z*ROC in our experiments are consistent with differences in the variances of summed similarity on *T* and *L* trials. Specifically, the summed-similarity values for *L* trials used in Experiment 3 had larger variance than did the values for *T* trials—a fact that signal detection theory predicts (and our simulations confirm) would produce *z*ROC slopes that are greater than 1. One take-home lesson may be obvious: By altering the variance of the summed similarity signals that would be generated on *T* and *L* trials, one can produce a wide range of *z*ROC slopes above and below a value of 1.0. Another take-home lesson may be less obvious. Beginning with stimulus characteristics, one can go directly to statements about the resulting distributions of summed-similarity signals on different kinds of trials.

GENERAL DISCUSSION

Learning From the Form of ROCs

For all three experiments reported in the present article, *z*ROC curves had slopes greater than 1. This was true in Experiments 1 and 2, in which recognition performance was estimated from order identification judgments, as well as in Experiment 3, in which recognition performance was measured directly—that is, from recognition judgments themselves. The results of our ROC analyses differ from those that have been reported for studies of verbal recognition memory. As was already discussed, we predicted and then found *z*ROC slopes of 1.1–1.3; studies using verbal materials report slopes of 0.8–1.0 (see, e.g.,

Donaldson & Murdock, 1968; Murdock, 1982; Ratcliff, McKoon, & Tindall, 1994; Yonelinas, 1997). What is the origin of this striking difference? Because our stimuli and our task both differ in several ways from those used to measure verbal recognition memory, it is impossible for one to know which factors are actually responsible for the divergent outcomes. For example, all three of our experiments produced monotonic positional gradients showing a pronounced recency effect, but no evidence of a primacy effect. These monotonic gradients, which were also seen in previous experiments using similar stimuli and tasks (Kahana & Sekuler, 2002; Yotsumoto et al., 2007), differ from a common finding with rehearsable stimuli—a relatively pronounced recency effect accompanied by some primacy effect. In addition to the very different nature of the stimuli in these two domains, each list in our study was followed by a single test probe (either a target or a lure), whereas studies of verbal episodic recognition use long sequences of targets and lures. Moreover, our study employed short lists of only three study items, whereas studies with verbal material have used considerably longer lists of study items (often on the order of 40–100 items). Our model-based prediction of *z*ROC slopes greater than 1.0 hinges on the relatively small variability in similarity that would be generated by comparing **p** and the one study item that was replicated by **p** (see above). In such an account, the number of lure items in the study list plays a crucial role. As that number grows, the contribution of the lone nonlure item will be diluted in comparison with the contribution of similarity signals from the increasing numbers of lure items. In turn, this dilution will decrease the differential in summed-similarity variability on *T* and *L* trials, producing *z*ROCs whose slopes approach a value

of 1.0. Note that this theory-based prediction does not lead to zROC slopes less than 1.0, as was reported by several investigators with lists of verbal items. In order to produce zROC slopes less than 1.0, an additional factor—such as variability in attention or goodness of encoding (see, e.g., Kahana, Rizzuto, & Schneider, 2005)—must be considered. Arguably, with verbal items, variability in goodness of encoding—particularly in long lists—would be greater than the variability associated with stimuli like those used in the present experiments. This additional variability could come from rehearsal (our stimuli are not easily rehearsable), variation in items' imagability, concreteness, word frequency, meaningfulness, and number of implicit associative responses (Hall, Sekuler, & Cushman, 1969), to name a few. All of these things make words interesting and complicated, but they probably add variability—to target items in particular. As Hintzman (1988) pointed out, differential variability in goodness-of-encoding for target items—especially across high- and low-frequency words—could account for key features of the mirror effect, a robust phenomenon that has resisted other explanations.[2] Whatever the origin of differences in ROC analyses may be, the combination of summed-similarity computations for recognition memory and a signal detection framework for decision making do a satisfactory job of connecting zROC slopes to the distributional characteristics of T and L trials in our experiments.

Sources of Item-Order Errors

The results displayed in Figure 5 suggest that structural and order similarity both influence misidentifications of serial position, and that their separate effects are approximately additive. In order to understand how structural similarity might lead to errors in order judgment, consider a summed-similarity account of false alarms in recognition judgments. Kahana and Sekuler's (2002) summed-similarity model assumes that three study items—$\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$—are stored in memory as noisy exemplars. When the probe—$\mathbf{p}$—is presented, the set of similarities between $\mathbf{p}$ and each of the noisy exemplars is computed (e.g., $\eta_1 \ldots \eta_3$). Again, subscripts signify the serial order of stimulus presentation. NEMo describes similarity values as exponentially decreasing functions of spatial differences between $\mathbf{p}$ and the corresponding values—$\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{m}_3$. From the resulting similarity values, the summed similarity—$\Sigma\eta$—is computed. In NEMo, a value of $\Sigma\eta > k$, where k is an optimal criterion, constitutes evidence that at least one of the study items matched $\mathbf{p}$, which makes $\mathbf{p}$ seem familiar. Over trials, the probability of a recognition response (that is, a *yes* response) corresponds to the proportion of trials on which $\Sigma\eta > k$.

The exemplar-similarity framework for understanding recognition memory can be extended to account for serial position judgments as well. Although we will not attempt to delineate a full account of such an extension, here is a sketch of some paths that such an extension might take.

A simple approach would be to base serial position judgments on the familiarity of items in memory. In NEMo, older items are coded with more noise and given less weight in the summed-similarity calculation (Ka-

hana et al., 2007). As such, probe items that have greater summed similarity to the list would be considered more recent. A problem with this account is that repeating an item should make it seem more recent, yet the data show that frequency and recency judgments are dissociable and possibly even independent (Flexser & Bower, 1974; Kahana & Loftus, 1999; McElree & Dosher, 1993).

One alternative approach assumes that positional information is coded as part of the memory trace of each studied item. This idea harks back at least to Ladd and Woodworth (1911), who assumed that positional information served as part of the cue for recall in serial learning tasks. More recently, positional or temporal coding has been an integral element in theories of serial recall, free recall, and even item recognition (Brown, Preece, & Hulme, 2000; Dennis & Humphreys, 2001; Howard & Kahana, 2002). Assuming that position information is encoded with each item, some mechanism is needed to read out, or retrieve this information. If, instead of summing the pairwise similarities between the probe and each of the list items, we retrieved the memorial representation of the studied item that was most similar to the probe, then the positional information encoded with that item could be used to drive judgments of serial position.

In this approach, each trial's pairwise similarities would be processed with a max operator, which returns the index of the largest item in η_1, η_2, or η_3. In the absence of error, this index would correspond to the serial position of the study item that most closely matches $\mathbf{p}$; this value could be the basis for serial-order judgments. Because of noise associated with each exemplar, there will be trials on which the index returned by max will correspond not to the serial position whose study item physically matched $\mathbf{p}$, but instead to the serial position of another, nonmatching study item. On such trials, the model would generate an item-order error, misidentifying the serial position of the matching item's similarity to $\mathbf{p}$. The probability of such errors would be some monotonically decreasing function of the $\mathbf{p}$'s similarity to study items occupying different serial positions. In other words, study items that were not replicated by $\mathbf{p}$ but were perceptually similar to it would be more likely misidentified as the match than would study items that were less similar to $\mathbf{p}$. This is the pattern of results shown in Figure 5.

A different mechanism is required to motivate the position-dependent misidentifications. Drawing on an account of analogous effects in free recall (Howard & Kahana, 1999, 2002), we assume that the representation in memory of each noisy exemplar is tagged with a position code. Because position tags can be degraded (partially forgotten) as a result of passage of time and/or interference, serially adjacent positions in a sequence would more likely be confused with one another than would positions more widely separated in a sequence. In the case at hand, loss of serial position information would make it more likely that $\mathbf{s}_1$ is misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_3$. Conversely, it would also make $\mathbf{s}_3$ more likely to be misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_1$. This account is mute on errors involving the forgetting of $\mathbf{s}_2$'s serial position. Again, this is the pattern of results seen in Figure 5. We should note that this ef-

fect is reminiscent of Dodson, Holland, and Shimamura's (1998) demonstration that even when subjects misidentify the source of information, they sometimes retain partial information about that source.

For terminological convenience, we have referred to the source of misidentifications just discussed as *positional*. However, we acknowledge that our data do not allow us to distinguish items' positional (order) information from various alternative formulations, such as time-based information. How the serial order of items is represented in short-term memory has been a topic of debate since Lashley (1951) formulated the issue and proposed his ingenious solution. The literature currently has almost as many models as it has data. Henson's (1998) review distinguishes among three classes of models for serial order in short-term memory: "chaining," "positional," and "ordinal," with positional models being subdivided into those that are temporal in nature (such as oscillator- or time-based models; Brown et al., 2000), those in which items are associated with their ordinal positions, taking no account of time (Burgess & Hitch, 1999; Howard & Kahana, 2002), and those in which an item's position is coded relative to a sequence's start and end (see, e.g., Henson, 1999). Any of these models could account for the near-neighbor position errors we observed, although each might take a different route to that end. A variety of experimental strategies could be used to distinguish among the competing accounts. These include methods that decouple our procedure's strict correlation between time-based and order-based stimulus information.

Finally, we should acknowledge potential limitations on the generalizability of our analysis of misidentifications. The first potential limitation arises from the length of our study lists. Each study list comprised just three items, two of which—the first and the last—might be considered special. Indeed, one account of memory for serial order explicitly asserts that the positions of items in a sequence are coded relative to the sequence's start and end (Henson, 1998, 1999). As a result, we cannot rule out the possibility that further experiments using substantially longer study lists would alter the proportion and/or distribution of identification errors from those observed here. A second potential limitation relates to the set of structural similarities embodied in our stimuli. Those similarities did influence overall recognition performance and were presumably responsible for some fraction of misidentification errors. One would expect that an overall increase or decrease in those structural similarities would not only alter overall recognition performance, but would also produce a corresponding change in the proportion of misidentifications that could be attributed to structural similarity.

On Choosing Stimuli for Studies of Memory

Similarity plays a strong role in the computations underlying episodic memory in a wide range of verbal tasks (see, e.g., Deese, 1959; Hintzman, 1988; Robinson, 1927; Roediger & McDermott, 1995; Skaggs, 1925). Recognizing this fact, we were encouraged to examine memory with stimuli whose similarity structure could be easily measured and controlled (Sekuler & Kahana, 2007). Al-

though gratings like the ones we used are certainly not the only stimulus class that satisfy these criteria, gratings are among the best understood in terms of their representation at various stages of the visual system. It would be interesting for researchers to determine whether our conclusions concerning ROC functions generalize to other classes of metric stimuli that have been used to study memory, including synthetic human faces (Yotsumoto et al., 2007), colors (Nosofsky & Kantner, 2006), and complex sounds (Visscher, Kaplan, Kahana, & Sekuler, 2007). An answer would reveal whether our results were unique to the particular perceptual stimuli whose representations in early visual are well characterized, or whether they could be generalized to other classes of stimuli.

REFERENCES

BROWN, G. D. A., NEATH, I., & CHATER, N. (2007). A temporal ratio model of memory. *Psychological Review*, **114**, 539-576.

BROWN, G. D. A., PREECE, T., & HULME, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, **107**, 127-181.

BURGESS, N., & HITCH, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, **106**, 551-581.

DEESE, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, **58**, 17-22.

DENNIS, S., & HUMPHREYS, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, **108**, 452-478.

DODSON, C., HOLLAND, P., & SHIMAMURA, A. (1998). On the recollection of specific- and partial-source information. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1121-1136.

DONALDSON, W., & MURDOCK, B. B. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology*, **76**, 325-330.

ESTES, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.

ESTES, W. K. (1994). *Classification and cognition.* Oxford: Oxford University Press.

FLEXSER, A. J., & BOWER, G. H. (1974). How frequency affects recency judgments: A model for recency discrimination. *Journal of Experimental Psychology*, **103**, 706-716.

GARNER, W. R., & HAKE, H. W. (1951). The amount of information in absolute judgements. *Psychological Review*, **58**, 446-459.

HALL, J. W., SEKULER, R., & CUSHMAN, W. (1969). Effects of IAR occurrence during learning on response time during subsequent recognition. *Journal of Experimental Psychology*, **79**, 39-42.

HENSON, R. N. A. (1998). Short-term memory for serial order: The start–end model. *Cognitive Psychology*, **36**, 73-137.

HENSON, R. N. A. (1999). Positional information in short-term memory: Relative or absolute? *Memory & Cognition*, **27**, 915-927.

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

HOWARD, M. W., & KAHANA, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 923-941.

HOWARD, M. W., & KAHANA, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, **46**, 269-299.

KAHANA, M. J., & LOFTUS, G. (1999). Response time versus accuracy

in human memory. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 322-384). Cambridge, MA: MIT Press.

KAHANA, M. J., RIZZUTO, D. S., & SCHNEIDER, A. R. (2005). Theoretical correlations and measuared correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 933-953.

KAHANA, M. J., & SEKULER, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research*, **42**, 2177-2192.

KAHANA, M. J., ZHOU, F., GELLER, A. S., & SEKULER, R. (2007). Lure similarity affects visual episodic recognition: Detailed tests of a noisy exemplar model. *Memory & Cognition*, **35**, 1222-1232.

LACROIX, J. P. W., MURRE, J. M. J., POSTMA, E. O., & VAN DEN HERIK, H. J. (2006). Modeling recognition memory using the similarity structure of natural input. *Cognitive Science*, **30**, 121-145.

LADD, G. T., & WOODWORTH, R. S. (1911). *Elements of physiological psychology: A treatise of the activities and nature of the mind from the physical and experimental point of view.* New York: Charles Scribner's Sons.

LAMBERTS, K., BROCKDORFF, N., & HEIT, E. (2003). Feature-sampling and random-walk models of individual–stimulus recognition. *Journal of Experimental Psychology: General*, **132**, 351-378.

LASHLEY, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112-136). New York: Wiley.

LEE, C. L., & ESTES, W. K. (1977). Order and position in primary memory for letter strings. *Journal of Verbal Learning & Verbal Behavior*, **16**, 395-418.

LOFTUS, G. R., & MASSON, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.

MACMILLAN, N. A., & CREELMAN, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

MALMBERG, K. J., & XU, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review*, **13**, 99-105.

MCELREE, B., & DOSHER, B. A. (1993). Serial recovery processes in the recovery of order information. *Journal of Experimental Psychology: General*, **122**, 291-315.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

MURDOCK, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, **89**, 609-626.

NACHMIAS, J., & STEINMAN, R. M. (1963). Study of absolute visual detection by the rating-scale method. *Journal of the Optical Society of America*, **53**, 1206-1213.

NOSOFSKY, R. M. (1986). Attention, similarity, and the identification categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

NOSOFSKY, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 363-394). Hillsdale, NJ: Erlbaum.

NOSOFSKY, R. M., & KANTNER, J. (2006). Exemplar similarity, study list homogeneity, and short-term perceptual recognition. *Memory & Cognition*, **34**, 112-124.

PAGE, M. P. A., & NORRIS, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, **105**, 761-781.

PELLI, D. G., ROBSON, J. G., & WILKINS, A. J. (1988). Designing a new letter chart for measuring contrast sensitivity. *Clinical Vision Sciences*, **2**, 187-199.

RATCLIFF, R., MCKOON, G., & TINDALL, M. (1994). Empirical generality of data from recognition memory ROC functions and implications for GMMs. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 763-785.

RATCLIFF, R., SHEU, C.-F., & GRONLUND, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, **99**, 518-535.

ROBINSON, E. S. (1927). The "similarity" factor in retroaction. *American Journal of Psychology*, **39**, 297-312.

ROEDIGER, H. L., III, & MCDERMOTT, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.

SEKULER, R., & KAHANA, M. J. (2007). A stimulus-oriented approach to memory. *Current Directions in Psychological Science*, **16**, 305-310.

SEKULER, R., KAHANA, M. J., MCLAUGHLIN, C., GOLOMB, J., & WINGFIELD, A. (2005). Preservation of episodic visual memory in aging. *Experimental Aging Research*, **31**, 1-12.

SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

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

SKAGGS, E. B. (1925). Further studies in retroactive inhibition. *Psychological Monographs*, **34**, 1-60.

STERNBERG, S. (1966). High-speed scanning in human memory. *Science*, **153**, 652-654.

VISSCHER, K., KAPLAN, E., KAHANA, M. J., & SEKULER, R. (2007). Auditory short-term memory behaves like visual short-term memory. *Public Library of Science Biology*, **5**, e56.

WATSON, C. S., RILLING, M. E., & BOURBON, W. T. (1964). Receiver-operating characteristics determined by a mechanical analog to the rating scale. *Journal of the Acoustical Society of America*, **36**, 283-288.

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

YONELINAS, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, **25**, 747-763.

YOTSUMOTO, Y., KAHANA, M. J., WILSON, H. R., & SEKULER, R. (2007). Recognition memory for realistic synthetic faces. *Memory & Cognition*, **35**, 1233-1244.

ZHOU, F., KAHANA, M. J., & SEKULER, R. (2004). Short-term episodic memory for visual textures: A roving probe gathers some memory. *Psychological Science*, **15**, 112-118.

NOTES

1. Unlike other summed-similarity models, NEMo posits an additional process by which the degree of summed similarity required to endorse a **p** item is modulated by the level of interitem similarity. However, since the inclusion of this mechanism does not alter the model's predictions of the zROC, we have presented the simpler special case in which β is absent.

2. The mirror effect is the empirical finding that with two classes of stimuli—say, high- and low-frequency words—the class that produces the higher hit rates also produces the lower false alarm rates.