Distinctiveness and the Recognition Mirror Effect: Evidence for an Item-Based Criterion Placement Heuristic

Ian G. Dobbins  
Duke University

Neal E. A. Kroll  
University of California, Davis

Superior detection and rejection of 1 versus another class of items during recognition is called the mirror effect. Some mirror effects may involve strategic adjustments based on item distinctiveness and its relation to memorability. Three experiments demonstrated mirror effects for known versus unknown scenes and 1 suggested a similar pattern for faces. In opposition to preexperimental familiarity, lures from known and frequently encountered locations were confidently rejected more often than unknown lures. Forgetting and speeding recognition reversed this lure response pattern, suggesting abandonment of strategic adjustment in favor of a single fixed criterion. With sufficient response time and recent encoding, observers demand more evidence for conceptually distinctive items, perhaps because such items typically foster vivid recollection during retrieval.

Keywords: recognition, mirror effect, distinctiveness, criterion, signal detection theory

One of the more heavily studied recognition memory phenomena is the mirror effect (for review see Glanzer et al., 1993). This effect can be observed between test lists when a strengthening manipulation (e.g., increased study duration) simultaneously improves target detection (hits) and lure rejection (correct rejections) in the strongly compared with the weakly encoded test lists. Mirror effects can also occur across two classes of stimuli within the same test list, which is the paradigm examined here. During these within-list mirror effects, one class of items is discriminated better than another, simultaneously yielding superior hit and correct rejection rates. The most studied within-list mirror effect occurs as a function of word frequency, with low-frequency words demonstrating higher hit and correct rejection rates than high-frequency counterparts. Because the mirror effect is frequently observed, it has been characterized as a "regularity of recognition" that may represent the operation of fundamental recognition mechanisms (Glanzer et al., 1993).

Current accounts of the mirror effect draw heavily upon signal detection theory (SDT; for review see Macmillan & Creelman, 1991). The basic signal detection model assumes that recognition decisions are based on continuous evidence values, which on average are higher for studied than for lure items, and which we refer to as familiarity or strength. The simplest SDT model assumes that the studied and lure item distributions are normal with equal variance. Critically, because the distributions overlap, observers are forced to use a cut off or criterion for responding "old." The response criterion itself can be based directly on each item’s familiarity, or it can be postulated to rely on an additional statistical computation of likelihood given that familiarity value. We consider the implications of our results for a likelihood criterion model in the discussion; however, we initially assume responding is based directly on the familiarity of each test item.

Under the signal detection model, within-list mirror effects are explained by either assuming that observers systematically adjust the response criterion for different classes of items (e.g., Hirshman, 1995) or that they instead use a single fixed-familiarity criterion, but the familiarity distributions of the item classes are arranged such that a mirror effect emerges. These different accounts have been labeled Type I and Type II respectively by Stretch and Wixted (1998). In order for a Type I model to be viable, the nominal class distinction (e.g., frequency) must map to an item characteristic that observers can reliably detect in order to shift the response criterion. Furthermore, such controlled switching presumably requires more time than simply maintaining a single fixed criterion throughout testing as it would require controlled executive processes.

Drawing on Stretch and Wixted (1998), the present Figure 1 illustrates five possible signal detection models. In all models, the class of items in the top of each panel, shown in solid lines (more memorable), is encoded better than is the class of items represented in the bottom of each panel, shown in dashed lines (less memorable). In addition, the left column illustrates models yielding a mirror effect across classes, whereas the right column illustrates concordant or partially concordant effects in which the increased hit rate of the more memorable class is accompanied by a reduced or similar correct rejection rate compared with less memorability.

---

1 It is also the case that if a more and less memorable class of stimuli are used in two different lists, then we would also expect a mirror pattern across lists.
memorable items. In other words, for the models in the right column, the encoding benefits conveyed to more memorable items do not translate into a rejection advantage for lures from this class. Models A and B assume that the lures from each class have the same preexperimental familiarity. The difference is that in Model A the observer strategically shifts the criterion upward for the more memorable class of stimuli; in contrast, in Model B the observer maintains a single fixed criterion throughout. Models C and D in Figure 1 show that the familiarity of lures for the more memorable items is assumed to be higher than that for the less memorable. Such a difference could arise as a result of preexperimental factors such as increased prior exposure. As with Models A and B, the only difference between Models C and D is that in the former observers adopt a more stringent criterion for the more memorable items, whereas in the latter a single criterion is used. Models A and C represent the construct of subjective memorability outlined by Brown et al. (1977; see also Benjamin & Bawa, 2004; Stretch & Wixted, 1998). A key premise of the subjective memorability account is that observers adopt a more stringent criterion for items deemed likely to be remembered if studied. In the context of word-frequency mirror effects, the subjective memorability account holds that observers view low-frequency words as more memorable and therefore use a more stringent criterion for this class of items. If one assumes that low-frequency targets receive superior encoding, then, because of the criterion shift, this model can predict a mirror pattern even if high- and low-frequency lures are assumed to have the same familiarity (Model A).

In terms of predicting the response proportions of hypothetical experiments, neither Model B nor Model D predict a mirror effect. In Model B, the correct rejection rates for the two classes should be similar, but the hit rate for memorable items should exceed that of the less memorable. In contrast, in Model D, more memorable items should have both a higher hit rate and a lower correct rejection rate due to the combination of increased memorability and preexperimental familiarity differences. We refer to Models B and D as nonstrategic because they do not require observers to systematically detect a characteristic of the item in order to shift the response criterion. In the case of the two criterion models, Models A and C, the predictions are more flexible and both are capable of producing a mirror effect because of the more stringent criterion used for the more memorable class. We refer to these two criterion models as strategic because they require the observer to make a trial-by-trial adjustment of the criterion. Furthermore, if observers were precluded from using this strategy, and instead reverted to using a single response criterion, Models A and C would reduce into the single criterion case represented to the right of each.

Model E (see Figure 1), illustrates what is necessary to model a mirror effect with a single fixed-response criterion. In Model E, the mirror effect results because the lures of more memorable items are assumed to be less familiar than are the lures of the less memorable class, an assumption typically made in dual-process models of the word-frequency mirror effect (e.g., Joordens & Hockley, 2000), which assume that the different false-alarm rates arise because of different levels of preexperimental exposure. It is important to note that in order for a single criterion model to transform from a mirror pattern (Model E) to a concordant pattern (e.g., Model D), the relative positions of the lure distributions must

---

2 The use of the term *concordant* here is potentially confusing as researchers typically report hits and false alarms. Under this convention, a simultaneous increase in the hit and false-alarm rate of a class of items is concordant. Because we instead list hits and correct rejections, which are the complement of false alarms, a concordant pattern is one in which the hit rate is increased and the correct rejection rate is decreased for a particular class of items.
be reversed while the relative positions of the target distributions remain the same.

Here we test the hypothesis that observers actively adjust the criterion within the test list in a manner consistent with either Model A or Model B. Under this subjective memorability hypothesis, we assume that the distinctiveness of the conceptual characteristics of each item, relative to those of other items in the test, affects the placement of the recognition criterion. The model we propose predicts that, under typical testing, observers will perceive conceptually distinct items as more memorable and therefore use a more stringent criterion for endorsing these items. Distinctiveness as a construct has been used many different ways; however, here we simply use it to connote the degree to which an item is associated with readily accessible conceptual knowledge that is not shared with other items in the test list. The benefit of such an assumption is that it is resistant to associative interference (i.e., a reduced “fan” effect; Anderson & Bower, 1974) and focusing on distinctive information promotes superior recognition (Hunt & Einstein, 1981). Our primary manipulations were focused on the difference between stimuli that were more likely to be identified as personally unique to the participants versus those that were more likely represented at a more superordinate level. This framing of conceptual distinctiveness is highly similar to the notion of differentiation, in which it is assumed that increasing exposure to items can lead to the acquisition of unique features or attributes in the representation of that item (Gibson & Gibson, 1955; McClelland & Chappell, 1998).

The subjective memorability model that we propose has three key aspects. First, we assumed that the encoding efficacy of an item is positively correlated with its conceptual distinctiveness. So, for example, seeing a picture of one’s favorite pizza restaurant would lead to superior encoding than would seeing a photo of an unknown pizza restaurant (despite the fact that both are clearly identified as pizza restaurants and are of an equal level of perceptual complexity). Second, for the types of items that we examined (scenes and faces), we also assumed that the conceptually distinctive items would have a higher preexperimental level of familiarity because the unique characteristics of the items have arisen through multiple prior encounters. Indeed a considerable amount of preexposure can yield item expertise, which has been shown to confer a reduced "fan" effect; Anderson & Bower, 1974) and focusing on distinctive information promotes superior recognition (Hunt & Einstein, 1981). Our primary manipulations were focused on the difference between stimuli that were more likely to be identified as personally unique to the participants versus those that were more likely represented at a more superordinate level. This framing of conceptual distinctiveness is highly similar to the notion of differentiation, in which it is assumed that increasing exposure to items can lead to the acquisition of unique features or attributes in the representation of that item (Gibson & Gibson, 1955; McClelland & Chappell, 1998).

The goal of this experiment was to examine whether a mirror effect could be observed in a stimulus class for which the lures of the more memorable class were assumed to be more preexperimentally familiar than were the lures of the less memorable class. This is opposite from what is typically assumed under dual-process models of the word-frequency mirror effect, which instead assume that the more memorable stimulus class has a lower, not higher, baseline familiarity (e.g., Joordens & Hockley, 2000; Reder et al., 2000). Such an outcome would be at odds with Models B and D in Figure 1 because neither suggests a greater ability to correctly reject lures from the more memorable stimulus class. Instead, mirror effects for these stimuli would suggest that observers are altering their criterion for responding on a trial-by-trial basis, supporting either Models A or C in Figure 1, or that the unknown lures were more familiar than the known lures and observers used a fixed criterion, Model E.

To further test the models, we also imposed a rapid response deadline for scenes and faces under the prediction that such speeding would force observers to abandon any potential criterion strategy because of time constraints and would yield a concordant pattern in the more memorable class of stimuli (known faces and locations) because of inherent differences in preexperimental familiarity. If speeding produced a concordant pattern (an increase in hit rates with a reduction in rejection rates for the known items), then the single criterion account would be suspect (Model E) because it would require the assumption that the lure distributions somehow reversed relative positions between the self-paced and speeded conditions. For the strategic models, the prediction was that speeding responding would force observers to use a single criterion. If speeded performance resembled Model D, then this would provide evidence that in the self-paced case Model C is correct. That is, Model C reduces to Model D when observers are forced to use a single criterion. Similarly, if speeded performance resembled Model B, then that would provide evidence that Model A was in fact the correct model in the self-paced case. In short, because speeding is assumed to force reliance upon a single criterion, it can be determined whether the preexperimental familiarity of the lures differs as a function of prior exposure, and, in turn, this can be used as evidence to infer the appropriate model in the self-paced case. Finally, using the scene stimuli, we also tested whether forgetting would lead to the abandonment of the proposed strategic criterion heuristic. Under the subjective memorability account, observers use the distinctiveness of the items to adjust the criterion because this predicts memorability. If, however, the relative memorability of the two classes is greatly reduced by forgetting, then observers are no longer predicted to use the heuristic.

**Experiment 1A, 1B, and 1C (Scenes)**

The goal of this experiment was to examine whether a mirror effect could be observed in a stimulus class for which the lures of the more memorable class were assumed to be more preexperimentally familiar than were the lures of the less memorable class. This is opposite from what is typically assumed under dual-process models of the word-frequency mirror effect, which instead assume that the more memorable stimulus class has a lower, not higher, baseline familiarity (e.g., Joordens & Hockley, 2000; Reder et al., 2000). Such an outcome would be problematic for Models B and D, which assume a single response criterion because neither predicts a mirror effect. A secondary goal was to extend within-list mirror effects to a class of stimuli other than words. Given that the effect has been predominantly researched through the use of verbal...
materials (e.g., word frequency, concreteness, word vs. nonword), it is important to demonstrate this effect in other stimuli, particularly those with high-ecological validity. To this end, we chose to use photographs of scenes likely known to the observers and scenes likely novel to the observers. If the subjective memorability model is correct, we should be able to observe a mirror effect favoring discrimination of known locations. The most important aspect of the predicted effect revolves around the correct rejection of known versus unknown locations. According to a subjective memorability model, and in opposition to the single criterion Models B and D, observers should be able to reject known lures more frequently and more confidently than unknown lures. Under the subjective memorability model, failure to recover evidence for an item judged memorable constitutes an additional form of information that facilitates correct rejections. Given two lures with similarly low signals, the one judged more subjectively memorable should be rejected more confidently. Thus even though one might assume that lures drawn from known locations will on average be more familiar than those drawn from unknown locations, the subjective memorability model suggests that known lures may nonetheless be rejected more often and more confidently. This would again run counter to Models B and D.

Method

Participants. Thirty-four undergraduate students from the University of California (UC), Davis participated in Experiments 1A and 1B either to fulfill course requirements or obtain extra credit, with the data from 1 participant removed due to uniformly poor memory performance in each study, resulting in 16 participants in each. In Experiment 1C, 16 undergraduate students from the University of California, Davis and 16 from Duke University participated in order to fulfill course requirements or for a remuneration of $10. Informed consent was obtained in accordance with Human Subjects Review Committees at both universities.

Materials and procedure (Experiments 1A and 1B). A total of 550 pictures were collected from the UC Davis campus, the surrounding area, and other more distant locations both in the United States and in foreign countries. On the basis of pilot ratings and experimenter judgments, we selected 200 for use assuming that approximately half were likely to be identified by the participants. Photographs were presented in color on a 17-in. monitor set at 800-pixel screen size.

The photographs were divided into two initial sets of 100, each having an equal number of likely to-be-known and likely to-be-unknown pictures and each serving as either targets or distractors equally often across 8 participants. Additionally, pictures were presented in four possible sequential orders for each study–test combination, under the constraint that no more than five lures or targets appear sequentially. Then, the 200 pictures were redivided into two sets and, in a similar fashion, assigned to an additional 8 participants. An additional 20 photographs were used as primacy and recency buffers during the study phase for all 16 participants.

Experiment 1A consisted of three phases: study, test, and knowledge rating. During the study phase, 120 pictures were displayed for 3 s each, and participants were instructed to try to remember each for an unspecified subsequent memory test.

Immediately following study, an old–new recognition test was administered in which participants rated their recognition memory on a confidence scale of 1 (most confident new) to 6 (most confident old) for studied and unstudied items. Before beginning, participants were urged to use the full range of ratings. The first 28 test items were not scored and constituted practice trials used to ensure participant understanding. These initial trials consisted of 14 new pictures and 14 of the primacy and recency buffers.

The remaining 200 test trials constituted the test proper. Accuracy of memory judgment was emphasized and responding was self-paced.

Following the final test trial, participants were again shown the critical 200 tested pictures (both previous targets and lures) and asked to indicate their personal experience, or lack thereof, with each photographed location prior to the experiment (knowledge rating phase). Participants indicated their knowledge of the location on a four point scale where 1 = have not seen the location before, 2 = may have seen the location a few times, 3 = have seen the location a large number of times, and 4 = have seen the location extremely often. Participants were told to base their ratings on the location represented in each picture and to ignore minor details such as the particular angle at which the picture was taken.

Results and Discussion

For the 100 studied photographs, posttest ratings indicated that on average 49 were rated as known (ratings of 3 or 4). For the lures, 45 were on average rated as known. Hits and correct rejection rates were conditionalized as a function of personal experience (known [rating of 3 or 4] vs. unknown [rating of 1 or 2]) and the data suggested a mirror effect (see Table 1). Using the signal detection measure $d'$, we found that participants were more accurate in discriminating known ($d' = 2.83$) than unknown ($d' = 1.85$) locations, $t(15) = 6.51, p < .001$. Planned comparisons for the hit and false-alarm rates demonstrated a higher hit rate for known versus unknown locations, $t(15) = 5.31, p < .001$, and suggested a higher correct rejection rate for known versus unknown locations, although this latter comparison only approached significance, $t(15) = 1.98, p = .066$. Inspection of the pattern for individuals showed that 15 of the 16 demonstrated the hit rate effect ($p < .01$, sign test), whereas 12 of 16 demonstrated the correct rejection pattern ($p = .080$, sign test).

In addition to response rates, we also considered the mean confidence during correct rejections, as these data were potentially relevant to distinguishing the models. Under SDT, confidence is monotonically related to the average distance of the stimuli from

$\text{3 We did not include the common signal detection estimate of bias (c) because it is ambiguous across conditions in which accuracy changes, if observers use a familiarity-based criterion (for review see Pastore et al., 2003). The bias measure c represents positive or negative deviation from the midpoint between the old and new item distributions and therefore is not fixed with respect to the familiarity scale (Macmillan & Creelman, 1991). Because of this, even if observers maintain a fixed familiarity criterion across two conditions, if those conditions yield different accuracies, then the estimate of c must change. For example, in the data of Morrell et al. (2002) differential strengthening of one category of items (e.g., professions vs. locations) yielded differences in the hit rates but not false-alarm rates from the strengthened category (Experiment 1: strong hit = .96, strong false alarm = .22; weak hit = .77, weak false alarm = .23). This pattern was consistent with the maintenance of a fixed familiarity criterion across the categories because the false-alarm rates were statistically equivalent. However, if one were to calculate the bias estimate c, it would necessarily change because the intersection of the distributions must change if accuracy changes across conditions. In the Morrell et al. data above, c changes from 0 during the weak to −.49 in the strong condition, but it would be erroneous to assume the criterion shifted unless one had evidence that observers reference their criterion to the intersection of the distributions. Instead, if observers base their criterion on a familiarity value, as assumed under the models considered here, then the false alarm–correct rejection rate (or its associated Z score) is a viable measure of criterion location across conditions and the one we use in the current article.}$
Experiment 1B

the response criterion. Thus, if familiarity values are equivalent or higher on average for known versus unknown lures, one would expect participants to be less confident when rejecting known lures provided a single criterion is used. When directly compared however, confidence was higher when rejecting known lures than unknown lures, \( t(15) = 6.51 \), \( p < .001 \). This pattern was observed in all 16 subjects (\( p < .001 \), sign test).

Although the rejection rate portion of the mirror effect did not quite reach significance, it is important to consider the nature of the lures. The data suggest that participants were (a) numerically more likely to reject as studied lures that they had previously encountered numerous times prior to testing, and (b) that they were more confident when doing so. This outcome is clearly discrepant with single criterion Model D because it predicts the reverse outcome, and it weighs against the single criterion Model B, although the nonsignificant \( t \) result suggests caution. Furthermore, the difference in confidence for the rejection of lures is also consistent with the subjective memorability account outlined in Brown et al. (1977).

Experiment 1B

The procedures of Experiment 1B closely follow those of Experiment 1A. However, we shifted the knowledge rating of the future target items from the test phase to the study phase for two reasons. First, we wondered whether having seen the items earlier during the study phase potentially contaminated the knowledge ratings of these items at test, similar to the false fame effect documented by Jacoby and colleagues (Jacoby et al., 1989). More important, we assumed that by directly focusing the participant’s attention toward whether or not he or she knew each item during study, we would increase the likelihood that participants would closely attend to the distinctive stimulus features and increase the potential size of any mirror-effect tied to this attribute (Hunt & Einstein, 1981; Hunt & Smith, 1996). Knowledge ratings for the lures were obtained by conducting a posttest rating task identical to that used in Experiment 1A. Note that this design results in two sets of knowledge ratings for the studied items (one during study and one following recognition testing).

Of the 100 photographs studied, posttest ratings indicated that on average 40 were rated as known (ratings of 3 or 4). For the lures, 37 were on average rated as known. Conditionalized hits and correct rejections demonstrated a mirror effect (see Table 1). Participants were more accurate in discriminating known (\( d’ = 2.89 \)) than unknown (\( d’ = 1.80 \)) locations, \( t(10) = 7.77, p < .001 \). The reduced degrees of freedom occurred because 4 participants had a perfect correct rejection rate and 1 had a perfect hit rate for known scene stimuli.\(^4\) Planned comparisons demonstrated a higher hit rate for known versus unknown locations, \( t(15) = 8.52, p < .001 \), and a higher correct rejection rate for known versus unknown locations, \( t(15) = 3.26, p < .01 \). All 16 individuals demonstrated the hit rate effect (\( p < .001 \), sign test), whereas 13 of 16 demonstrated the correct rejection effect (\( p = .024 \), sign test). Although confidence was numerically higher for correct rejections of known versus unknown lures, the difference was not statistically significant, \( t(15) = 1.72, p > .10 \).

Combined with Experiment 1A, these data demonstrate that a within-list mirror effect can be obtained with scenes that likely differ only with respect to personal experience. This effectively rules out the single criterion Models B and D (shown in Figure 1), at least under conditions of immediate testing and self-paced responding, and is consistent with the hypothesis that participants actively adjust the criterion during testing (Models A and C). To our knowledge, this is the first demonstration of a mirror effect with scene stimuli, and it is noteworthy given the high-ecological validity of these stimuli. Nonetheless, Model E can also potentially accommodate the data by using a single response criterion. For that model to hold in the current study, one has to assume that lures from unknown locations are more, not less, familiar than known locations. One potential way this might occur, suggested by a reviewer, is that the unknown locations may somehow be more prototypical and hence elicit greater familiarity than the known locations. There are two potential objections to this suggestion. First, under exemplar models of recognition one might instead assume that frequently visited locations (i.e., known) would have greater representation in terms of stored exemplars, and thus one’s favorite pizza restaurant would likely yield more not less memory evidence than a randomly selected photo of a pizza restaurant (Hintzman, 1986). Second, under Model E, it is not clear how observers were able to reliably rate the known and unknown locations following study. If the unknown locations were more prototypical, and this led to an increased sense of familiarity, then it is unclear how participants were able to distinguish known from unknown locations during the ratings. Despite this, it nonetheless could be the case that there was a systematic difference in the structural–perceptual qualities of the known and unknown items that led to the observed mirror effects. For example, through sampling error, known scenes could have been more perceptually detailed or vivid than unknown scenes and this could be responsible for the increased discriminability. Experiment 1C further examines this possibility.

\(^4\) Although there are potential corrections for \( d’ \) when faced with floor or ceiling response rates (Macmillan & Creelman, 1991), we chose not to apply these because \( d’ \) estimates are known to be unreliable at extreme response rates. However, given that instances of perfect hit or rejection rates occurred exclusively for known stimuli, omitting these scores simply made the comparisons somewhat more conservative and did not change the outcome of these comparisons.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Hits M SD</th>
<th>Correct rejections (CR) M SD</th>
<th>d’</th>
<th>CR confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>.88 .08</td>
<td>.91 .09</td>
<td>2.83</td>
<td>2.55</td>
</tr>
<tr>
<td>Unknown</td>
<td>.70 .09</td>
<td>.87 .09</td>
<td>1.85</td>
<td>2.22</td>
</tr>
</tbody>
</table>

Table 1
Results From Experiments 1A and 1B
Experiment 1C

Although the data from Experiments 1A and 1B clearly suggest a mirror effect as a function of preexperimental location knowledge, one might be concerned that the result may somehow reflect a systematic physical–structural difference between pictures rated as known versus unknown. To control for this possibility, we conducted a study across two campuses in which each campus served as the control for the other. A second concern that this study addressed is whether the knowledge ratings themselves somehow distorted our findings. The primary purpose of these ratings was to appropriately sort items because the exact same set of items would not be known for each individual. Experiment 1C allowed us to determine whether the effect still obtained in a less powerful design in which ratings were not used to determine known and unknown locations for each participant. To address these concerns, we tested UC Davis and Duke University students by using pictures from both campuses, under the prediction that the mirror effects would be reversed across each campus because the semantic distinctiveness of the items would be based on participant experience and not the structural characteristics of the stimuli.

The methods of Experiment 1C were simplified in comparison to Experiments 1A and 1B. During study, participants passively viewed 45 pictures from the UC Davis surroundings, 45 pictures from the Duke surroundings, and 48 pictures taken from other sources (24 well-known and 24 obscure—these 48 pictures were included just to make the list longer and were not included in the subsequent memory test). These three sources of pictures were quasi-randomly interspersed throughout the study list with different orderings used across participants. In addition, there were 12 pictures (three Davis, three Duke, three famous, and three obscure) used for primacy and recency buffers. During study, each picture was presented for 2.5 s in preparation for an unspecified memory test. Immediately following study, participants were given an old–new recognition test consisting of the 45 Davis and 45 Duke studied items, together with an equal number of matched lures. In addition, the first 32 trials of the test constituted practice trials constructed from 16 of the original primacy and recency buffers and 16 new items. During test, incorrect responses were followed by feedback indicating when an error was made. All responses were made with the keyboard and responding was self-paced.

The data were subjected to a Participant Group (Duke vs. UC Davis) × Picture Origin (Duke vs. UC Davis) × Response (hit or correct rejection) three-way mixed analysis of variance (ANOVA). The ANOVA yielded a main effect of picture origin, \( F(1, 30) = 20.25, p < .001 \), suggesting that Davis photos were more discriminable than were Duke photos, and a main effect of response, \( F(1, 30) = 6.82, p < .05 \), indicating more correct rejections than hits. In addition, there was a two-way Participant Group × Picture Origin crossover interaction, \( F(1, 30) = 84.61, p < .001 \), that resulted because the Duke students responded correctly more often to Duke photos (.85) than to Davis photos (.80), \( F(1, 15) = 8.88, p < .01 \); whereas the Davis students responded correctly more often to Davis photos (.92) than to Duke photos (.77), \( F(1, 15) = 123.84, p < .001 \). Finally, there was a three-way interaction between all three variables, \( F(1, 30) = 13.45, p < .001 \) (see Figure 2).

As illustrated in Figure 2, the three-way interaction occurred because the pattern of hit and correct rejection rates is largely reversed across campuses. Planned comparisons confirmed this impression, showing a hit rate advantage for the home campus photos for Duke students (.81 vs. .72), \( t(15) = 4.27, p < .001 \), and Davis students (.92 vs. .74), \( t(15) = 8.18, p < .001 \). Duke students failed to demonstrate a home campus photo advantage for correct rejection rates, however there was a home campus advantage in correct rejection rates for Davis students (.91 vs. .81), \( t(15) = 5.38, p < .001 \).

The data demonstrate that the mirror patterns observed in Experiments 1A and 1B were not driven by systematic differences in the perceptual characteristics of the stimuli. If that were the case,

Figure 2. Results of Experiment 1C. The data demonstrated a qualitatively different mirror pattern for University of California, Davis and Duke University students as a function of stimulus origin. Whereas Davis students demonstrated a mirror effect favoring discrimination of Davis photos (greater hit and correct rejection rates), Duke students tended to display the reverse, showing superior performance for discriminating Duke relative to Davis photos (greater hit rates), \( F(1, 30) = 13.45, p < .001 \).
the same pattern should have emerged for Duke students; that is, superior hit and correct rejection rates for UC Davis stimuli. Instead, Duke students showed a largely reversed pattern in comparison to Davis students. Overall, the mirror pattern in the Duke students was less robust compared with the pattern in the Davis students, but this is not surprising given that extensive piloting was used in the selection of the Davis pictures ensuring that a high proportion would in fact be known to the participants. In contrast, the Duke pictures were selected as likely to be known but were not thoroughly screened using pilot ratings, and it is likely that a smaller proportion were in fact uniquely identifiable and hence distinctive to the students. Nonetheless, the data confirm that the mirror pattern emerges because of experience dependent knowledge regarding the stimuli, not because of systematic differences in their physical construction or inherent typicality (i.e., perceptual level effects). The data also suggest that the rating procedure and conditionalized scores used in Experiments 1A and 1B are not the source of the effect.

Experiment 2A and 2B: Speeded Response

In total, Experiment 1 suggests that a single criterion model of performance appears inappropriate for these materials; although one can endorse the single criterion Model E if willing to assume that lures from unknown scenes are perceived as more familiar than known ones. Turning to the two criterion models (Models A and C), the patterns observed across the studies in Experiment 1 are equally consistent with both because either one can accommodate a mirror effect and the confidence data observed. The fundamental difference between these models is whether or not there is a preexperimental difference in the familiarity of the lures. In order to distinguish among these models, Experiment 2 used a speeded response technique. Assuming that systematically adjusting the criterion requires controlled processing, we anticipated that sufficiently speeding performance would force the participants to instead apply the same fixed criterion to all stimuli. If there is a preexperimental familiarity difference, such that known lure locations are perceived as more familiar, then a correct rejection advantage should emerge for the unknown stimuli (Model D). If instead the lures have similar familiarity regardless of whether they are known or unknown, then the correct rejection rate would be similar (Model B).

Method

Participants. Twenty undergraduate students from the University of California, Davis participated in Experiment 2A to fulfill course requirements. Data from two were discarded because of inconsistent preexperimental familiarity ratings (>24% change) and 2 others were discarded as a result of near chance recognition performance. In Experiment 2B, a larger number of participants were collected because we assumed that speeded responses would introduce greater variability into the results. Thirty-five undergraduates participated in Experiment 2B to fulfill course requirements. Data from 2 were discarded because of inconsistent familiarity ratings and data from 1 other were discarded because of poor recognition performance. Informed consent was obtained in accordance with Human Subjects Review Committees at the University of California, Davis.

Materials and procedure. The materials and study phase of Experiment 2 were identical to those of Experiment 1B, however when participants failed to rate the item for knowledge within 3 s during the study phase, the screen remained blank until a response was entered. On average this occurred very rarely, representing less than 5% of the study trials. In the test phase, participants made a simple binary old–new discrimination of the pictures instead of the confidence scale rating. For Experiment 2A recognition was self-paced, whereas during Experiment 2B participants were given 500 ms to respond. If they failed to respond within this time limit, the trial was discarded. This occurred on 18.0% of the 240 test trials. The first 12 trials were practice trials constructed of new items and primacy or recency buffers.

Results and Discussion

Experiment 2A: Self-Paced Recognition

Of the 100 photographs studied, ratings indicated that on average 54 were rated as known (ratings of 3 or 4), whereas for the lures, 52 were on average rated as known. The means for hits and correct rejections as a function of personal experience (known vs. unknown location) did not demonstrate a full mirror pattern in the self-paced data (see Table 2). As indexed by $d'$, participants were more accurate in discriminating known ($d' = 3.03$) versus unknown ($d' = 1.91$) locations, $t(15) = 7.49, p < .001$. Planned comparisons demonstrated a higher hit rate for known versus unknown locations, $t(15) = 9.98, p < .001$, however there was no statistical difference between the correct rejection rates of known and unknown locations. Inspection of the pattern for individuals showed that all 16 demonstrated the hit rate effect ($p < .001$, sign test).

The failure to find a mirror pattern in the correct rejection data must be weighed against the results of Experiment 1, which indicate a small but reliable advantage for the known lures. Further, the equivalence is difficult to accommodate under a single criterion model unless the lures are assumed to have the same preexperimental familiarity across the known versus the unknown classes. The results of Experiment 2B suggest that this is unlikely.

Experiment 2B: Speeded Recognition

Of the 100 studied targets, on average 48 were rated as known (ratings of 3 or 4), whereas for the lures, 39 were on average rated as known. The means for hits and correct rejections as a function of personal experience (known vs. unknown location) did not demonstrate a mirror pattern, instead they clearly demonstrated a concordant pattern (see Table 2). Participants remained more accurate in discriminating known ($d' = 1.77$) versus unknown ($d' = 1.30$)

<table>
<thead>
<tr>
<th>Response</th>
<th>Scene</th>
<th>Hits $M$</th>
<th>Hits $SD$</th>
<th>Correct rejections $M$</th>
<th>Correct rejections $SD$</th>
<th>$d'$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 2a</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-paced</td>
<td>Known</td>
<td>.94</td>
<td>.05</td>
<td>.87</td>
<td>.11</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>.75</td>
<td>.10</td>
<td>.87</td>
<td>.06</td>
<td>1.91</td>
</tr>
<tr>
<td><strong>Experiment 2b</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speeded</td>
<td>Known</td>
<td>.90</td>
<td>.08</td>
<td>.64</td>
<td>.16</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>.68</td>
<td>.14</td>
<td>.76</td>
<td>.13</td>
<td>1.30</td>
</tr>
</tbody>
</table>

Table 2

Results of Experiments 2A and 2B
1.30) locations, \( t(31) = 4.84, p < .001 \). Planned comparisons demonstrated a higher hit rate for known versus unknown locations, \( t(31) = 8.19, p < .001 \). In contrast to the prior data demonstrating a correct rejection advantage for known lures, or equivalent rejection rates for known and unknown lures, correct rejection rates were greater for unknown versus known lures, \( t(31) = 6.74, p < .001 \). For individuals, 31 of 32 demonstrated the hit rate effect (\( p < .001 \), sign test), and 28 of 32 demonstrated correct rejection rates indicating greater familiarity for the known versus unknown lures (Model D; \( p < .001 \), sign test).

These data complement the findings of Experiment 1 and suggest that the preexperimental familiarity of known scenes is higher than that for unknown scenes. Thus, overall it appears that the mirror effects observed in Experiment 1 can be more securely attributed to Model C in Figure 1 because this is the two criterion model that assumes that known lures are more familiar than unknown lures during self-paced responding but that this difference is opposed by a systematic criterion adjustment. During self-paced responding, Model A is ruled out because it predicts similar rejection rates during speeded responding (Model B). The data are also problematic for the single fixed criterion Model E. For this model to remain viable, not only must one assume that unknown lures are more familiar than known lures during self-paced responding, one must further assume that this relationship reverses when responding is speeded (Model D). We know of no mechanism predicting this reversal within a familiarity-based criterion framework.

**Experiment 3A and 3B: Retention**

Experiments 3A and 3B were designed to examine a second prediction of the subjective memorability model, namely, that participants would abandon it if distinctiveness no longer clearly correlated with greater memorability during testing. To this end, we inserted a 1-week retention interval between study and test, which was based on research indicating such delays result in a decline in subjective experiences of remembering (Gardiner & Java, 1991) and on the prediction that the delay would greatly reduce, but not eliminate, the memory advantage for the known class. The primary prediction was that in contrast to Experiment 1 and similar to speeded performance in Experiment 2, the correct rejection rate would be higher for unknown versus known stimuli; as with speeding, this would imply that the heuristic pictured in Model C of Figure 1 had been abandoned in favor of Model D. Furthermore, despite the predicted reversal of correct rejection rates across delayed versus immediate testing, we assumed that known stimuli would remain somewhat more discriminable than unknown stimuli.

**Method**

**Participants.** Seventeen undergraduate students from the University of California, Davis participated in Experiment 3A to fulfill course requirements. Data from 1 participant were discarded because of a lower consistency across the study rating and posttest rating phases (32% of the targets received a different rating following testing). In Experiment 3B, a larger number of participants were collected because we assumed that the 1-week delay would introduce greater variability into the results. Forty-one undergraduates participated in Experiment 3B to fulfill course requirements. Data from 9 were discarded. Seven participants demonstrated poor consistency in the familiarity ratings (24% or greater change in ratings for targets), and the remaining 2 failed to return for the second session. Informed consent was obtained in accordance with Human Subjects Review Committees at the University of California, Davis.

**Materials and procedure.** We used 240 pictures as possible study and test stimuli, with 16 possible sequences created for counterbalancing similar to Experiments 1A and 1B. As in Experiment 1B, there were three sequential parts of Experiment 3A: study, test, and familiarity rating. During the study phase, participants were informed that a long list of pictures would be presented and that their task was to indicate their level of preexperimental familiarity with each picture by using the same 4-point frequency scale as in Experiment 1A and 1B. Each picture was presented for 3 s, after which it disappeared from the screen. If a participant failed to respond within 3 s, the screen remained blank until a response was entered. In addition, they were instructed to rate the pictures carefully for the following two reasons: (a) a later part of the experiment would require rating of the same pictures again, and (b) ratings across the two parts would be compared for consistency. Twenty-four pictures, 12 at the beginning and 12 at the end, served as a practice trial. During the test phase, confidence-based old–new recognition discriminations were required identical to those in Experiments 1A and 1B. No more than four targets or lures were presented in sequential order and 32 pictures, comprised of 16 new pictures and 16 buffer items, were presented at the beginning of the test phase to serve as practice trials. Responding was self-paced.

In the final phase, all previous pictures were rated for preexperimental familiarity on the same 4-point scale. The only difference between Experiment 3A and 3B was the length of the retention interval. Whereas testing was immediate in Experiment 3A, there was a 1-week interval between study and test for Experiment 3B.

**Results and Discussion**

**Experiment 3A: Immediate Recognition**

Of the 100 studied targets, 48 on average were rated as known, whereas for the lures, 47 on average were rated as known. The means demonstrated a mirror effect (see Table 2), with participants more accurate in discriminating known (\( d' = 3.31 \)) than unknown (\( d' = 2.08 \)) locations, \( t(11) = 7.70, p < .001 \). The lower degrees of freedom resulted because of perfect scores in the known condition for either hits or false alarms which precluded the calculation of \( d' \). Planned comparisons demonstrated a higher hit rate for known versus unknown locations, \( t(15) = 5.95, p < .001 \), and higher correct rejection rates for known versus unknown locations, \( t(15) = 5.17, p < .001 \). All 16 participants demonstrated the hit rate effect (\( p < .001 \), sign test), and 15 of 16 demonstrated the correct rejection effect (\( p < .01 \), sign test with one tie). Additionally, participants were more confident when correctly rejecting known locations than unknown locations, \( t(15) = 2.31, p < .05 \).

**Experiment 3B: Delayed Recognition**

Of the 100 studied targets, 51 on average were rated as known whereas for the lures, 53 on average were rated as known. Interspersing 1-week delay between study and test eliminated the mirror effect, instead resulting in a concordant pattern (see Table 3). Participants remained more accurate in discriminating known (\( d' = 1.60 \)) than unknown (\( d' = 1.33 \)) locations, \( t(31) = 3.08, p < .01 \), despite the increased retention interval. Planned comparisons demonstrated a higher hit rate for known versus unknown locations, \( t(31) = 11.63, p < .001 \); however, now the correct rejection
rate for unknown locations clearly exceeded that of known locations, \(t(31) = 7.02, p < .001\), which is the reverse of what was observed for immediate testing. The pattern for individuals showed that 31 of 32 demonstrated the hit rate effect \((p < .001, \text{sign test})\), and 30 of 32 demonstrated the reversed rejection rate effect with respect to Experiment 3A \((p < .001, \text{sign test})\). Finally, confidence did not differ for the correct rejection of known and unknown locations \((t < 1)\).

The results of Experiment 3 again suggest a criterion placement heuristic consistent with Model C in Figure 1 that appears to be abandoned for Model D when substantial forgetting occurs and the memorability of known and unknown items does not greatly differ. Thus following the 1-week delay, the rejection rate appears to be heavily affected by the preexperimental familiarity of the material with respect to a single criterion. Again, as with Experiment 2, these data rule out Models A and B and are also difficult to reconcile under the notion that observers always maintain a single fixed criterion (Model E). Under this latter account, one would have to assume that forgetting leads to a reversal in the perceived familiarity of known and unknown lure distributions. Again, we know of no familiarity-based forgetting model that predicts this.

**Experiment 4A and 4B: Faces**

The above experiments suggest that participants apply a different criterion to known versus unknown locations when given ample response time following minimal forgetting. Our working hypothesis is that this occurs because the unique conceptual information that is typical of known locations is, under normal circumstances, associated with the recovery of memorable encoding experiences. If correct, this account should generalize to other stimulus classes that potentially contain item specific, distinctive information. The class we tested in Experiments 4A and 4B were famous faces. More specifically, we anticipated that although the baseline familiarity of famous faces is higher than nonfamous faces, this would be partially countered by the adoption of a more stringent response criterion for the famous faces during self-paced testing. Furthermore, as famous faces are associated with distinctive conceptual information, we anticipated that encoding would lead to greater memorability for this stimulus class. Finally, as with scenes, we predicted that speeding responses would prevent the heuristic, and therefore rejection rates for the unknown class would exceed those of the known class, eliminating any potential mirror pattern.

**Method**

**Participants.** Twenty-four undergraduate students from the University of California, Davis participated in each experiment to fulfill course requirements. Informed consent was obtained in accordance with Human Subjects Review Committees at the University of California, Davis.

**Materials and procedure.** On the basis of pilot study data, 352 pictures of faces were chosen such that each face had some chance of being recognized as famous by our student population. These faces were drawn from the Internet and consisted of movie stars (both those of present day and those whose films are frequently shown on TV), politicians, historical figures, popular sports figures, TV personalities, rock stars, and so forth. There was considerable variation among the students concerning which faces were identifiable. That is, some students were much more likely to recognize the faces of politicians than those of movie stars, others were more likely to recognize the faces of rock stars, and so forth. Using the rankings obtained in a pilot study, we assigned the faces to two subsets in such a way as to maximize the probability that participants would recognize approximately the same number of faces in each subset. An equal number of participants received each subset as a target list and as a lure list. Eight different sequences of presentation were used across participants. The screen resolution was set at 640 pixels \(\times 480\) pixels and was black except for a white 162-pixel \(\times 152\)-pixel window with a red frame within which the faces, ranging from 110 pixels \(\times 150\) pixels to 160 pixels \(\times 150\) pixels in size, were displayed. Each face was presented in black and white with only the face and the hair showing.

During study, each participant was told that the experiment was a test of visual memory, that a long series of faces would be shown at a very fast rate, and that they would then be given a subsequent memory test. The study phase preceding self-paced (Experiment 4A) and speeded (Experiment 4B) tests were identical. Each participant encoded a sequence of 176 faces, with each face shown for approximately 600 ms followed by 200 ms blank interval before the next face.

During both tests, the first 72 test trials were made up of the first 18 and the last 18 faces of the study trials along with 36 lures. These trials were not included in the analysis in order to (a) minimize potential primacy or recency effects, (b) provide a filled retention interval for the remaining items, and (c) provide a practice set (which was particularly useful for the speeded testing group of Experiment 4B). On the basis of pilot data, faces of differing likelihood of being known were interleaved during both the study and test sequences.

In Experiment 4A, participants were given the same confidence rating instructions used previously, whereas in Experiment 4B they were given

<table>
<thead>
<tr>
<th>Interval</th>
<th>Scene</th>
<th>Hits</th>
<th>Correct rejections (CR)</th>
<th>CR confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Immediate</td>
<td>Known</td>
<td>.96</td>
<td>.04</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>.81</td>
<td>.12</td>
<td>.89</td>
</tr>
<tr>
<td>1-week delay</td>
<td>Known</td>
<td>.87</td>
<td>.08</td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>.61</td>
<td>.14</td>
<td>.83</td>
</tr>
</tbody>
</table>

| Experiment 3A |           | M    | SD                     | M  | SD  | d'  | CR confidence |
|---------------|-----------|------|------------------------| M  | SD  | d'  | CR confidence |
| Immediate     | Known     | .96  | .04                    | .95 | .05  | 3.31 | 2.90          |
|               | Unknown   | .81  | .12                    | .89 | .06  | 2.21 | 2.82          |
| 1-week delay  | Known     | .87  | .08                    | .62 | .20  | 1.60 | 2.33          |
|               | Unknown   | .61  | .14                    | .83 | .12  | 1.33 | 2.36          |
the same speeded response instructions used earlier. In the case of speeded responding, during the first 72 practice trials, the time that the face was presented systematically decreased from 1,800 ms to 920 ms, which was the presentation time for the critical 280 trials. If the participant responded while the face was still on the screen, the participant’s response was portrayed on the screen for 100 ms. If the participant did not respond during the presentation of the face, the face was replaced with the words Too Slow! presented for 150 ms. This occurred on 7% of the 280 critical test trials.

During posttest knowledge ratings, each of the 352 faces were presented one face at a time and the participant attempted to tell the experimenter the name of the person being represented and the reason why that person was known (e.g., “Brett Favre, quarterback for the Green Bay Packers,” or “Laura Bush, wife of the current president”). The only data that were used in the analyses were those associated with the faces that were judged unknown (neither name nor conceptual information available) and those judged known (both name and conceptual information available). The data associated with faces that (a) the participant could name but not describe (there were very few of these), (b) describe but not name, or (c) misidentified, were discarded.

Results and Discussion

Experiment 4A: Faces, Self-Paced Recognition

Of the 140 photographs studied, on average 42 were rated as known, whereas for the lures on average 39 were rated as known. Although $d'$ indicated that participants were more able to discriminate among known ($d' = 1.71$) than among unknown faces ($d' = .56), $t(21) = 8.46, p < .001$; this advantage was restricted to hit rates which were notably higher for known than for unknown faces, $t(23) = 10.95, p < .001$ (see Table 4). Correct rejection rates for known versus unknown faces were numerically equivalent. Inspection of the pattern for individuals showed that all 24 demonstrated the hit rate effect ($p < .001$, sign test). Although the rejection rates were equivalent, participants were more confident when correctly rejecting known than unknown faces, $t(23) = 4.48, p < .001$.

Experiment 4B: Faces, Speeded Recognition

Of the 140 photographs studied, on average 56 were rated as known, whereas for the lures on average 51 were rated as known. Participants remained more able to discriminate between known ($M = 1.08$) than unknown faces ($M = .26$), $t(22) = 9.51, p < .001$, during speeded responding. Planned comparisons demonstrated that the hit rate was higher for known than for unknown faces, $t(23) = 17.86, p < .001$. In contrast, the correct rejection rate was higher for unknown stimuli, $t(23) = 5.28, p < .001$. All 24 participants demonstrated the hit rate effect ($p < .001$, sign test), and 22 of 24 demonstrated higher correct rejection rates for unknown versus known faces ($p < .001$, sign test).

Although the performance for unknown faces was very low, overall, the results with faces are similar to those observed with scenes. This suggests that the hypothesized criterion heuristic generalizes across stimulus types, provided that the known items are associated with distinctive, item-specific information. In addition, because the hypotheses deal mainly with the pattern of correct rejections under known and unknown conditions, the outcomes seem robust with respect to low-performance issues. During self-paced responding, participants demonstrated an advantage for the known stimuli in hit rates, with equivalent rejection rates for both known and unknown items. However, when responding was speeded, the rejection rate for the unknown stimuli exceeded that of the known stimuli, despite preservation of the hit rate advantage for known items. Assuming that participants use a single criterion during speeded responding, this indicates that the baseline familiarity of unknown faces is lower than that of known faces, making these items easier to reject during speeded responding.

General Discussion

The current research yielded two important and interrelated findings. First, to our knowledge, this is the first demonstration that known versus unknown scenes yield mirror effects. To the extent that current models do not anticipate these data, then one must either conclude that the mechanisms underlying mirror effects are highly specific to the materials used (cf. Greene, 2004) or that the models may require modification in order to generalize. Second, the data provide evidence that observers actively adjust the response criterion throughout the test list when time permits and when conceptual distinctiveness predicts memorability (Model C). Critically however, if speeded (Experiment 2) or substantial forgetting (Experiment 3) occurs, then mirror effects are not observed and this suggests use of a simpler, single criterion decision rule. As we discuss below, this finding is somewhat surprising given evidence that observers often appear reluctant to actively

Table 4

<table>
<thead>
<tr>
<th>Response</th>
<th>Face</th>
<th>$M$</th>
<th>$SD$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$d'$</th>
<th>CR confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 4A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immediate</td>
<td>Known</td>
<td>.79</td>
<td>.14</td>
<td>.78</td>
<td>.16</td>
<td>1.71</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>.39</td>
<td>.21</td>
<td>.78</td>
<td>.14</td>
<td>.56</td>
<td>2.20</td>
</tr>
<tr>
<td>Experiment 4B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speeded</td>
<td>Known</td>
<td>.71</td>
<td>.13</td>
<td>.67</td>
<td>.17</td>
<td>1.08</td>
<td></td>
</tr>
</tbody>
</table>
shift criterion during testing. First, however, we consider the relationship between distinctiveness and preexperimental familiarity.

Relation Between Distinctiveness and Preexperimental Familiarity

The fact that the distinctiveness of the current materials coincided with increased preexperimental exposure allowed the speeding and forgetting manipulations to demonstrate experimental dissociations in the correct rejections rates, and indeed this was the motivating factor governing our selection of these materials. However, the relationship between familiarity and item distinctiveness is typically assumed to be reversed under dual-process accounts of the word-frequency mirror effect (e.g., Joordens & Hockley, 2000; e.g., Reder et al., 2000). In the source of activation confusion model of Reder and colleagues, the word-frequency mirror effect is explained through a dual-process model consisting of contextual recollection and item familiarity. The advantage for low-frequency targets results because these items have fewer competing prior contextual associations compared with high-frequency targets, and this results in facilitated association with episodic content and hence greater recollection (i.e., less interference or competition; Reder et al., 2000; see also Cary & Reder, 2003; Gardiner & Java, 1990; Gutten tag & Carroll, 1997). However, the false alarm portion of the effect, when present, is typically assumed to be driven entirely by the different preexperimental exposures of high- and low-frequency lures, such that infrequently encountered lures are perceived as less familiar and hence endorsed less often than are their high-frequency counterparts. Thus as currently framed, dual-process accounts would not predict mirror effects with the stimuli used here because they do not assume that participants shift familiarity-based response criteria on a trial-by-trial basis.

Context of Memorability Estimates

A large obstacle to the subjective memorability account was a finding by Wixted (1992) that participants failed to explicitly appreciate that low-frequency words were in fact more memorable than their high-frequency counterparts. In fact using multiple rating paradigms, Wixted found that participants consistently and mistakenly rated high-frequency words as more memorable in hypothetical recognition situations, which is the reverse of what would be needed to generate a mirror pattern. However, this discrepancy has been potentially resolved in recent studies examining memorability judgments in the context of actual recognition testing, instead of during hypothetical memory situations as was the case in the Wixted study. When questioned regarding the memorability of items during actual testing, participants appear to reverse their beliefs regarding memorability of high- and low-frequency words and appropriately rate low-frequency words as more memorable during recognition (Benjamin, 2003; Gutten tag & Carroll, 1998). Thus, it is plausible that participants can detect and use item-specific characteristics diagnostic of memorability during testing.

Repet ted Failures to Observe Strategic Criterion Shifts

Recent studies specifically designed to induce participants to shift response criterion within a list on the basis of differential strengthening of one class of stimuli compared with another have failed to show within-list criterion shifts (Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998). For example, Stretch and Wixted (1998) had participants study high-frequency words five times (strong) but low-frequency words only once (weak), followed by testing in which high- and low-frequency words (both targets and lures) were each shown in two different colors. Under these conditions, participants should be able to adjust their criterion according to the anticipated strength of each class. That is, seeing the color and corpus frequency associated with strong encoding should lead the participants to expect a strong memory signal and hence use a conservative criterion, whereas the color associated with weak rehearsal should lead to the expectancy of a weak signal and application of a more lenient criterion. It is important to note that this should have led to a reversal or minimization of the typical word frequency mirror effect. However, although the hit rate portion was reversed with participants endorsing the highly rehearsed high-frequency items more often, the false alarm portion remained as predicted under a fixed criterion model; more false alarms were committed for high- versus low-frequency items (Experiment 3) and this general pattern was replicated in several cases. More recently, similar manipulations that used even more notable class differences (e.g., different semantic categories assigned to strong and weak rehearsal conditions) have also failed to provide evidence that the criterion is adjusted systematically during the test (Morrell et al., 2002).

Currently, we can only speculate as to why the criterion may be more labile here than in previous studies. First, as noted by Stretch and Wixted (1998), participants may often not be inclined to do the extra cognitive work necessary to make adjustments under many circumstances. In order to use a criterion heuristic in previous studies, participants would have had to note that one category was studied more often than another and to understand that they could adjust the criterion to maximize performance on the basis of this list-wide characteristic. However, to the extent that participants explicitly parse items into two distinct categories, they are also likely aware that during test each category contains an equal number of targets and lures. To the extent that they try to evenly distribute responses when uncertain, this may make them resistant to “biasing” their reports on the basis of this characteristic.

In contrast, consistent with the model of Brown et al. (1977), we suggest that assessments of memorability more naturally arise on a trial-by-trial basis in response the individual characteristics of each test item. It is important to emphasize that although we have referred to distinctive and nondistin ctive (or known and unknown) items by using a class nomenclature (and researchers often do the same when describing low- and high-frequency words), we assume that the individual items are “known” on the basis of information that tends to be different and uniquely assessed across items. For example, one’s favorite pizza restaurant, the registrar’s office, and the local ballpark may all be “known,” but this is not because they all share a common feature or features. Given this, we do not assume that participants shift their criterion because they identify items as arising from the ad hoc category “known,” but instead because these items tend to have unique features that participants assume they would have previously considered during study, and in turn expect to remember these prior reflections during testing.
The Nature of the Response Criterion

In the current study, we assumed that participants use a familiarity or strength-of-evidence criterion for each item. However, mirror effects have also been interpreted as relying on a fixed likelihood decision criterion that is derived from the original evidence values (e.g., Glanzer et al., 1993) and applied similarly across the more and less memorable stimulus classes. The likelihood value itself arises from the comparison of probability density values associated with the target and lure distributions for a given class of item (e.g., low-frequency words) and familiarity value. For example, if participants maintained the criterion at the intersection of the target and lure distributions for known and unknown items, this would correspond to a likelihood ratio of 1, and mirror effects would occur provided this likelihood value was maintained across known and unknown items in the current study (for a review, see Macmillan & Creelman, 1991). More specifically, Models A, C, and E in Figure 1 could be reinterpreted as demonstrating a mirror effect resulting from the maintenance of a constant likelihood ratio criterion of 1. There are at least two reasons we have instead chosen a familiarity- or evidence-based interpretation. First, the data of Stretch and Wixted (1998) and Morrell et al. (2002) do not favor the likelihood account for reasons noted above, namely that the false-alarm rate did not change across manipulations of memorability. Second, the results of speeding and forgetting here also argue against a fixed likelihood ratio account because under this account the correct rejection rate should always favor the more memorable stimulus class. That is, concordant patterns should not have been observed across known and unknown items because they significantly differ in discriminability. Instead, during speeding and forgetting in Experiments 2B, 3B and 4B, participants had higher correct rejection rates for the unknown items despite the fact that performance remained superior for known items (see also Benjamin & Bawa, 2004; Greene, 2004; see also Joordens & Hockley, 2000).

Similar to previous studies (Morrell et al., 2002; Stretch & Wixted, 1998), we have relied on comparisons of the correct rejection rates (or false-alarm rates) for different stimulus classes in order to infer something about the relative locations of their corresponding familiarity distributions. However, one objection raised to this approach by a proponent of the single strength-of-evidence criterion model (Models E and D) is that the correct rejection rates of the lures are not diagnostic of familiarity in any absolute sense, but instead reflect the outcome of two fundamentally separate discrimination processes between the targets and lures for the known and unknown stimulus classes (with the outcomes governed by within-class feature similarity). From this perspective, the rejection rates across known and unknown items, or across any dichotomous stimulus classification for that matter, cannot be meaningfully compared because they arise from two different discrimination problems that yield differently scaled evidence values. The validity of this criticism rests on two nested assumptions. The first is whether known and unknown items constitute two fundamentally different categories of items that are processed in isolation. As we note above, similar to the distinction between high- and low-frequency words, we do not assume that known and unknown items constitute a natural dichotomous category, and therefore the rejection rates (provided the criterion is fixed) can be used to infer the relative position of the lure distributions.

However, we agree that if the memory discrimination of known items occurs in isolation from, or is based on different evidence than, unknown items, then the observed response rates are not direct indicators of relative familiarity in a general or common sense. However, adopting this new assumption leads to an additional problem if one simultaneously maintains that the endorsements for both item classes are made with respect to same criterion and scale (Models E and D). If the discriminations for the two classes are made on the basis of evidence that is not directly comparable, how is this evidence then mapped onto the common strength-of-evidence criterion and its associated scale? One historical approach taken to avoid such relative scaling problems when dealing with potentially different types of evidence values has been to assume that the evidence values from each class are transformed into a common metric such as statistical likelihood. However, as noted above, the data do not support the likelihood account. Given this, we suggest that if one assumes a common and fixed strength-of-evidence criterion, then it becomes necessary to explain how the evidence is mapped onto this scale and why the ordinal positions of the lure distributions reverse with respect to the common criterion when transitioning from Model E to Model D. That is, we contend that this flexibility in the ordinal locations of the distributions needs to be explained as it constitutes an additional free parameter when accommodating data. In contrast to the fixed criterion account, the model we propose (Model C and Model D) assumes that the known lures are always perceived as more familiar than the unknown lures. Furthermore, this ordinal relation holds across forgetting or speeding and is only masked to the extent that participants are able to use subjective memorability to selectively adjust the response criterion. One prediction arising from this model is that participants with problems in executive control, such as frontal lobe patients (Stuss & Knight, 2002), may in fact demonstrate concordant response patterns with the current materials, despite good encoding and immediate testing, because they may fail to implement the subjective memorability heuristic.

References

New Editor Appointed, 2007–2012

The Publications and Communications (P&C) Board of the American Psychological Association announces the appointment of a new editor for a 6-year term beginning in 2007. As of January 1, 2006, manuscripts should be directed as follows:

- Emotion (www.apa.org/journals/emo.html), Elizabeth A. Phelps, PhD, Department of Psychology, New York University, 6 Washington Place, Room 863, New York, NY 10003.

Electronic manuscript submission. As of January 1, 2006, manuscripts should be submitted electronically via the journal’s Manuscript Submission Portal (see the Web site listed above). Authors who are unable to do so should correspond with the editor’s office about alternatives.

Manuscript submission patterns make the precise date of completion of the 2006 volumes uncertain. The current editors, Richard J. Davidson, PhD, and Klaus R. Scherer, PhD, will receive and consider manuscripts through December 31, 2005. Should 2006 volumes be completed before that date, manuscripts will be redirected to the new editor for consideration in 2007 volume.