Confidence–Accuracy Inversions in Scene Recognition: A Remember–Know Analysis

Ian G. Dobbins, Neal E. A. Kroll, and Qiang Liu
University of California, Davis

S. E. Clark (1997) offered a modified signal-detection explanation of the confidence–accuracy inversions observed in E. Tulving’s (1981) experiments. In addition to replicating E. Tulving (1981), we had participants make “remember–familiar” judgments. Confidence and accuracy dissociated across subjective reports. Response confidence differed only for judgments based on familiarity, whereas accuracy differed only for “remember” responses. S. E. Clark’s model does not predict this, nor can it mimic “remember” performance across all conditions. We propose that although “knowing” can be accommodated within an equal variance signal-detection account, “remembering” is governed by contextual constraints that influence the distinctiveness of information upon which participants rely during reports. The current paradigm is a pictorial analogue to H. L. Roediger and K. B. McDermott’s paradigm (1995) in that participants claim to explicitly remember thematically related items that were not actually seen during study.

In everyday situations, people often attempt to convince themselves and others of the accuracy of a claimed memory through expressions of confidence. In fact, even when it is apparent that memory is inaccurate, strong feelings of confidence often lead them to believe otherwise. Statements such as “I’m sure I left those keys on the dining room table; somebody must have moved them.” exemplify this tendency to defend the accuracy of confident memories. The reason for this tendency seems relatively clear; typically, confidence and accuracy are directly related (see Murdock & Dufty, 1972; Robinson & Johnson, 1996). Not only are people cognizant of this relation, they are notably perturbed when the two are not concordant. However, applied research regarding eyewitness testimony (for review, see Luus & Wells, 1994) and basic experimental paradigms (e.g., Chandler, 1994; Tulving, 1981) have demonstrated that confidence and accuracy can be reliably dissociated and even inversely related, with participants expressing greater confidence under conditions where they are less accurate.

Recently, these confidence–accuracy inversions have been taken as evidence against a simple signal-detection-based strength or familiarity account of scene recognition (Chandler, 1994). If both accuracy and confidence are assumed to rely on a single underlying trace strength or familiarity dimension, they should always be positively related. Experimental manipulations that improve accuracy presumably do so by strengthening target items, thus increasing the distance between target- and distractor-item distributions. If confidence is monotonically related to this distance, one would also expect some increase in confidence to accompany the increase in accuracy; in no case should confidence actually diminish.

Clark (1997) proposed an alternative signal-detection account of Tulving’s paradigm (1981) that suggests that confidence–accuracy inversions are attributable to differences in the variance of the difference distribution underlying two-alternative forced-choice (2AFC) responding. We discuss this account more fully following the presentation of the procedure and results of Tulving (1981).

Tulving (1981)

Participants were briefly shown a long series of scenic photographs (160) that were constructed by selecting either the right or left half of an entire picture. These items were designated A, B, C, D, and so forth. Following study, a 2AFC recognition test was administered with three experimental conditions. In condition A–A’, the original target is paired with the other half of the photo from which it was drawn. In this case, not only is the distractor (A’) similar to an item in memory (the original target presented during the study), but it is also similar to the concurrently presented target. Tulving (1981) designated these two types of similarity as ephoric and perceptual, respectively. The second condition of interest was designated A–B’. In this case, the distractor (B’) is the other half of a photo previously studied; however, it is not from the same scene as the current target and is therefore perceptually dissimilar. Finally, as a control condition (A–X), distractors were drawn from photos that the participant had not previously seen and thus were both perceptually and ephorically or mnemonicly dissimilar to anything previ-
ously studied. Following each discrimination, participants rated their selection confidence on a scale of 1 to 4, with 4 representing their greatest confidence. The results are presented in Table 1 and demonstrate the confidence-inversion previously discussed. The A–A’ condition is significantly more accurate than the A–B’ condition, yet the pattern in confidence is reversed. This pattern would seem incompatible with the standard signal-detection account if both accuracy and confidence were positively related to item strength.

A Signal-to-Noise Ratio Account

In accordance with signal-detection theory, Clark (1997) proposed that discrimination in this task is accomplished as a function of the difference in target and distractor strengths. Because selection in the task is dependent on the difference of target- and distractor-item familiarity distributions, the shape of the resulting difference distribution may be heavily influenced by whether target and distractor familiarities can be assumed to be independent on a trial-by-trial basis. The overall mean difference in strength for the A–A’ and A–B’ conditions is equivalent. In each condition, the target is an item that has been seen once, and the distractor is new but similar to an item in memory. However, the variance of a difference in variables or distributions is equal to sum of the variances less two times the covariance of the variables:

\[ \text{Var}[F(T) - F(D)] = \text{Var}[F(T)] + \text{Var}[F(D)] - 2\text{Cov}[F(T), F(D)], \]

where \text{Var} = variance, \( F(T) \) and \( F(D) \) = the probability density functions for the target and distractor strengths, respectively, \( T \) = target, \( D \) = distractor, and \text{Cov} = covariance. It is this variance component that enables the prediction of confidence-accuracy inversion because of the difference in the covariance of target and distractor familiarities in the A–A’ and A–B’ conditions. Because the items in the A–A’ condition are selected from the same scene, if one item is highly familiar, there will be a tendency for the other item to be highly familiar also. This situation is not the case for the A–B’ condition, in which familiarities should be independent. Thus, if performance is characterized by the signal-to-noise ratio \( F(T) - F(D)/\text{Var}[F(T) - F(D)] \), the denominator will be larger for the A–B’ condition than for the A–A’ condition. Figure 1 shows the resulting hypothetical difference distributions for the experiment.

![Figure 1. Strength-based model of discrimination performance based on the theory of Clark (1997). The figure shows the difference distribution \( F(T) - F(D) \) for all three conditions with relative mean strengths and relative variances illustrated. \( T \) = target; \( D \) = distractor; A–A’ = scenically and mnemonically similar distractor; A–B’ = scenically dissimilar but mnemonically similar distractor; A–X = novel distractor.](image)

All distributions lie to the right of zero on the strength dimension because on average targets are more familiar than distractors. The A–A’ and A–B’ distributions are centered on one another because on average they have the same difference in strength. The differences in width result because of the variance issues just discussed. Because of this variance difference, a greater proportion of the A–A’ condition lies to the right of the zero-difference strength marker. Therefore, one would predict greater accuracy in this condition. However, it can also be seen that on average, strength values for the A–B’ condition lie further away on the strength dimension from the marker than for the A–A’ condition. If confidence is monotonically related to strength, the A–B’ condition should display greater confidence. Thus, what is predicted overall is a confidence-accuracy inversion (see Clark, 1997, for confirming monte carlo results using MINERVA 2). Clark (1997) argued that this signal-to-noise model, instantiated in MINERVA 2 and several other global memory models, provides a simple account of the data without recourse to decision-level proposals.

In the present study, we explore a proposal that draws on discussions provided by Tulving (1981) and Chandler (1994), whose paradigm parallels that of Tulving’s in many ways. Tulving (1981) suggested that “highly similar test items may induce participants to engage in deeper or more elaborate processing of retrieval information (Craik & Jacoby, 1979), or examine the relevant evidence more thoroughly (Bower, 1972, p. 98) . . . with lower degrees of perceptual similarity, the probability that the distractor is (incorrectly) selected may be mainly determined by its ecphoric similarity” (p. 495). Similarly, Chandler (1994) suggested that distractor items related on a scene or general thematic level to previously seen items (i.e., ecphorically similar) will sometimes be incorrectly endorsed and may lead to the illusion of accurate remembering. Overall, this

### Table 1

**Confidence-Accuracy Inversion, Tulving (1981)**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Proportion correct</th>
<th>Hits</th>
<th>False alarms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A–A’</td>
<td>.74</td>
<td>2.15</td>
<td>1.52</td>
</tr>
<tr>
<td>A–B’</td>
<td>.68</td>
<td>2.46</td>
<td>2.11</td>
</tr>
</tbody>
</table>

Note. A–A’ = scenically and mnemonically similar distractor; A–B’ = scenically dissimilar but mnemonically similar distractor.
account holds that the accuracy difference between A–A′ and A–B′ conditions results because participants are lured into selecting distractors that closely resemble a previous event, and the resemblance is such that they will often falsely remember having seen the item. In the A–A′ condition, this tendency is somehow checked by the presentation of items from the same scene. Given the solid assumption that participants realize the photos are from the same scene, target selection presumably must occur at a more rigorous, or elemental, level. We refer to this explanation as the levels-of-selection account.

The goal of this current study is to compare these accounts by replicating the procedure of Tulving (1981) with one important modification. In the current study, participants were required not only to select the item previously encountered but also to report on whether the item was selected on the basis of recollective or nonrecollective information by using the remember–know paradigm (Tulving, 1985). Numerous variables have been demonstrated to dissociate remembering and knowing (for reviews, see Gardiner & Java, 1993; Rajaram & Roediger, 1997), and the dual-process model of recognition memory crafted by Jacoby and colleagues (e.g., Jacoby, 1991; Jacoby, Yonelinas, & Jennings, 1997; Yonelinas, 1994) has been theorized to tap the same underlying processes.

Within the dual-process account, participants are assumed to make recognition decisions based on at least two sources of information: recollection and familiarity. Recollection, leading to “remember” responses, is assumed to rely on the retrieval of qualitative information regarding the previous occurrence of an item. Familiarity, on the other hand, is presumed to rely on a non-specific feeling of prior occurrence, such that people know they have previously seen an item. Similar experiences occur often, for example, when one sees a person who strikes one as extremely familiar, but whose identity or occasion of previous acquaintance cannot be retrieved (Mandler, 1980). Recent dual-process theorists (e.g., Jacoby, 1991) have stressed the importance of recollection in the strategic control of responding. Given that the levels-of-selection account suggests that the accuracy difference between A–A′ and A–B′ conditions results because of participant-guided response strategies, it is natural to predict that such strategies will only be apparent when participants report having access to recollected information (i.e., during remembering).

In light of the levels-of-selection account, we expected high levels of false “remember” responses in the A–B′ condition compared with the A–A′ condition in which participants are forced to use a more rigorous or elemental level of selection for “remember” reports. In the absence of the contribution of recollection, participants are presumed to simply select which item seems most familiar. A recent dual-process model of recognition (Yonelinas, 1994) contends that memory based solely on familiarity adheres to an equal variance signal-detection process. As discussed earlier, this situation would lead to similar accuracy between A–A′ and A–B′ conditions because the targets have equal memory strength in both conditions. In the case of confidence, predictions are much less clear. Typically, “remember” responses in remember–know paradigms are accompanied by high degrees of confidence; however, it is not clear how the level of iconic information relied on would affect confidence based on recollection. For example, one might remember seeing a scene of a living room or may remember seeing a particularly ugly lamp or piece of furniture in that scene. Whether one memory should be more confident than the other is unclear. The case for confidence based on familiarity is also equivocal. Although strength considerations clearly influence confidence, numerous other decision-level factors may also influence expressed confidence.

With regard to the signal-detection account offered by Clark (1997), no clear distinction is made between qualitative and continuous sources of memory, and, therefore, one would expect confidence–accuracy inversions regardless of whether participants report that their selection was based on feelings of familiarity or recollective experience.

Method

Participants

Sixty undergraduates from the University of California, Davis participated in return for course credit.

Stimuli

From an initial pool of over 400 CD-ROM-based color photographs, a set of 190 pictures was selected on the basis of similarity ratings of the halves of each bisected picture. The pictures consisted of numerous scenes of natural and man-made locations from around the globe. Similarity ratings were conducted on a scale of 1 (not similar) to 5 (perfectly similar) by 10 participants not involved in the experiment proper. The items selected ranged in similarity from 4.1 to 3.4. During the study, 128 items were presented, with an additional 32 making up the pool from which novel distractors were selected for the A–X condition (see the Design and Procedure section). The remaining 30 were used as a set of fixed buffer items (10 for primacy and 20 for recency). The pool of 160 experimental items was counterbalanced across cell conditions and six possible sequences.

Design and Procedure

The current design closely followed that of Tulving (1981). After a brief description of the experiment, a short practice session was administered. Participants were shown a series of 8 practice items shown centrally on the computer monitor for 2 s each. They were then shown 6 practice-test pairs in a 2AFC format, with each picture half occupying the left and right side of a standard 14-in. monitor. Again, the three possible types of distractor conditions present were explained. In the A–A′ condition, the distractor was drawn from the same scene as the previously studied target and was therefore scenically similar to the correct alternative and similar to the memory for that alternative. In the A–B′ condition, the distractor was the other half of an item previously studied but not tested. It was therefore scenically dissimilar to the current correct alternative but similar to the memory of a previously studied item. Finally, in the A–X condition, the distractor was novel, neither resembling the current alternative nor another item previously
studied. Participants selected items by using the keyboard to enter the number corresponding to the confidence for the chosen item, 1–4 for the left item and 5–8 for the right. Following this step, a prompt appeared below the selected item, questioning whether it was chosen because the participant remembered seeing it (or an element within it) or because it seemed more familiar than the other item. Participants were instructed to respond “remember” (with the R key) whenever they recollected seeing the scene or particular elements in the scene during the study phase. If recollection was absent, they were instructed to respond “familiar” (with the F key). We chose to use the word “familiar” because students often confuse the more standard “know” response with an expression of high confidence. Following this step, the screen was blanked, and the next test trial began. Responses were self-paced. The experiment proper consisted of a series of 128 study items shown for 2 s each, immediately followed by 2AFC testing with 32 items in each of the conditions described above. Average completion time was approximately 45 min.

Results

For ease of interpretation, the data analysis was conducted in three parts. First, one-way analyses of variance (ANOVA) were conducted on the dependent variables of interest (accuracy, confidence for correct and incorrect responses) across the three distractor conditions (A–A’, A–B’, and A–X). Following this step, the data were broken down into “remember-familiar” responses, and the proportion of total “remember” responses was analyzed as a function of distractor condition. Finally, “remember-familiar” responses for the conditions of interest A–A’ and A–B’) were contrasted for accuracy (percentage correct) and for response confidence. All post hoc comparisons were done with Tukey’s honestly significant difference (HSD) tests with an alpha of .05 unless otherwise noted.

Overall Performance

A repeated measures ANOVA revealed a significant effect of distractor condition on accuracy scores (percentage correct), F(2, 118) = 53.08, MSE = .007, p < .001. Post hoc comparisons revealed that the A–X condition (M = 0.81) was more accurate than the A–A’ condition (M = 0.70), which was in turn more accurate than the A–B’ condition (M = 0.66, see Table 2).

A repeated measures ANOVA revealed a significant effect of distractor condition on confidence for correct responses, F(2, 118) = 31.15, MSE = .006, p < .001. Participants were less confident when selecting A–A’ targets (M = 2.68) than when selecting A–B’ (M = 2.98) or A–X (M = 2.97) targets for which confidence did not differ. Additionally, there was an effect of distractor condition on confidence for incorrect responses, F(2, 116) = 43.50, MSE = .014, p < .001. Participants were most confident when committing false alarms during the A–B’ condition (M = 2.59) compared with the A–A’ (M = 2.01) and A–X (M = 2.03) conditions, for which confidence did not differ.

These results replicate Tulving’s (1981) and demonstrate a clear dissociation of response confidence and accuracy during this task. Although participants were more accurate during the A–A’ than the A–B’ condition, they responded less confidently. Furthermore, there was also a difference in confidence when committing errors, with participants being significantly more confident of their errors in the A–B’ condition than in the A–A’ condition. Thus, as in Tulving’s (1981) study, the results show that recognition accuracy in this task was higher with scenically similar than with dissimilar distractors, whereas the same manipulation had the opposite effect on response confidence.

Remember-Familiar Analysis

The purpose of this analysis was to investigate whether the above differences between A–A’ and A–B’ conditions were independent of the participant’s introspective judgments regarding the basis of memory performance. The first one-way ANOVA indicated whether the overall proportion of “remember” responses differed as a function of distractor condition. Following this step, two-way ANOVAs were performed that crossed response type (“remember” or
with distractor condition (A–A' or A–B') for accuracy and response confidence.

A one-way repeated measures ANOVA, performed on the proportion of “remember” responses participants gave as a function of distractor condition, yielded a significant effect, \( F(2, 118) = 47.34, MSE = 0.008, p < .001 \). Participants gave “remember” responses significantly more often in the A–B' (\( M = .52 \)) and A–X (\( M = .49 \)) conditions than in the A–A' (\( M = .37 \)) condition. When considered in conjunction with the data above, this result suggests that although participants are less accurate in the A–B' than the A–A' condition, they are more likely to report having remembered seeing the chosen item. To further investigate this effect, two-way repeated measures ANOVAs with independent variables of distractor condition (A–A' vs. A–B') and response type (“remember” or “familiar”) were conducted on percentage correct scores. This metric is inadequate in yes–no designs because it may be heavily influenced by the participant’s tendency toward negative or positive responding (see Macmillan & Creelman, 1991; Swets, 1986); however, it is appropriate for 2AFC performance.

Results yielded significant main effects for distractor condition, \( F(1, 59) = 23.50, MSE = 0.015, p < .001 \), and response type, \( F(1, 59) = 192.20, MSE = 0.015, p < .001 \). An important result was that there was a Distractor Condition \( \times \) Response Type interaction, \( F(1, 59) = 29.17, MSE = 0.009, p < .001 \). Post hoc dependent \( t \) tests demonstrated that accuracy did not differ between the conditions when responses were made on the basis of feelings of familiarity (\( t < 1 \)). However, “remember” responses were more accurate in the A–A' condition (\( M = .88 \)) than in the A–B' condition (\( M = .74 \)), \( t(59) = 7.36, p < .001 \). Thus, the difference noted in accuracy for the A–A' and A–B' conditions above is carried exclusively by decisions accompanied by recollective experience. When participants responded solely on the basis of familiarity, the scenic similarity of the distractor did not influence the accuracy of responding. Figure 2 shows the receiver operating characteristics (ROCs) for “remember,” “familiar,” and overall responding (see Macmillan & Creelman, 1991, for logic and construction of ROCs) and demonstrates that the above accuracy effects are consistent across all levels of reported confidence. Frequency counts underlying the curves in Figure 2 are given in Table 3.

One important point regarding the differences in the ROCs for “remember” responses relates to the different tendencies of participants to use this response in the different conditions. The earlier analysis on total proportion of “remember” responses showed that the during A–B' and A–X conditions, participants were equally likely to use the “remember” response, whereas during the A–A' condition, they were significantly less likely to do so. On the basis of this observation, one might be tempted to characterize A–A' performance as simply more conservative than A–B'. However, the notion of conservative responding has different implications in 2AFC versus yes–no paradigms. In the latter, participants are assumed to adopt an absolute response criterion or boundary (e.g., yes–no, or know–remember) above which a given individual item will trigger a response. To be more conservative in the yes–no case simply means that this boundary is shifted upwards and that overall, the participant is less willing to make a positive response, regardless of whether the item is a target or distractor. Consequently, the accuracy measure (\( d' \)) is expressly independent of such shifts. In the case of 2AFC, however, responses are not made on the basis of an absolute item strength, but the difference in strength between target and distractor (see Figure 1). In this instance, to respond more conservatively would be to require a greater difference in strength before giving a particular response type (e.g., “remember”). Because this difference in strength is positively related to accuracy, the placement of any criterion for responding will be related to accuracy. To say someone is more conservative in Condition A versus Condition B simply restates the observation that they require more evidence in A than in B but does not address the source of this improved selectivity. Thus, conservative responding in the yes–no case simply represents an indiscriminate re-

![Figure 2](image-url)  
**Figure 2.** Receiver operating characteristics demonstrating cumulative accuracy as function of response confidence for overall, “remember,” and “familiar” responding (see Table 3). \( P = \) percentage; A–X = novel distractor; A–A' = scenically and mnemonically similar distractor; A–B' = scenically dissimilar but mnemonically similar distractor.
Table 3
Frequency Counts for Construction of Receiver Operating Characters (ROCs)

<table>
<thead>
<tr>
<th>Condition and target</th>
<th>Confidence</th>
<th>Left response</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Right response</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>245</td>
<td>125</td>
<td>137</td>
<td>174</td>
<td>115</td>
<td>90</td>
<td>52</td>
<td>38</td>
<td>976</td>
</tr>
<tr>
<td>A-A'</td>
<td>Left</td>
<td></td>
<td>25</td>
<td>55</td>
<td>73</td>
<td>129</td>
<td>154</td>
<td>119</td>
<td>141</td>
<td>248</td>
<td>944</td>
</tr>
<tr>
<td></td>
<td>Right</td>
<td></td>
<td>297</td>
<td>136</td>
<td>78</td>
<td>86</td>
<td>84</td>
<td>71</td>
<td>78</td>
<td>112</td>
<td>942</td>
</tr>
<tr>
<td>A-B'</td>
<td>Left</td>
<td></td>
<td>81</td>
<td>80</td>
<td>65</td>
<td>74</td>
<td>105</td>
<td>134</td>
<td>134</td>
<td>305</td>
<td>978</td>
</tr>
<tr>
<td></td>
<td>Right</td>
<td></td>
<td>400</td>
<td>174</td>
<td>137</td>
<td>121</td>
<td>74</td>
<td>49</td>
<td>42</td>
<td>22</td>
<td>1019</td>
</tr>
<tr>
<td>A-X</td>
<td>Left</td>
<td></td>
<td>22</td>
<td>33</td>
<td>51</td>
<td>70</td>
<td>125</td>
<td>122</td>
<td>141</td>
<td>337</td>
<td>901</td>
</tr>
</tbody>
</table>

Familiar

| Condition and target | Confidence | Left response | 21 | 73 | 117 | 172 | 113 | 83 | 38 | 13 | 630 |
|                      |            | Right response | 11 | 37 | 64 | 128 | 151 | 107 | 81 | 14 | 593 |
| A-A'                 | Left       |              | 15 | 77 | 73 | 79 | 77 | 61 | 45 | 7 | 434 |
|                      | Right      |              | 7 | 45 | 60 | 74 | 99 | 107 | 68 | 11 | 471 |
| A-B'                 | Left       |              | 28 | 94 | 119 | 121 | 73 | 49 | 29 | 8 | 521 |
|                      | Right      |              | 5 | 27 | 49 | 68 | 112 | 108 | 84 | 16 | 469 |

Remember

| Condition and target | Confidence | Left response | 224 | 52 | 20 | 2 | 7 | 14 | 25 | 346 |
|                      |            | Right response | 14 | 18 | 9 | 1 | 3 | 12 | 60 | 234 | 351 |
| A-A'                 | Left       |              | 282 | 59 | 5 | 7 | 7 | 10 | 33 | 105 | 508 |
|                      | Right      |              | 74 | 35 | 5 | 0 | 6 | 27 | 66 | 294 | 507 |
| A-B'                 | Left       |              | 372 | 80 | 18 | 0 | 1 | 0 | 13 | 14 | 498 |
|                      | Right      |              | 17 | 6 | 2 | 2 | 13 | 14 | 57 | 321 | 432 |

Note. ROCs are constructed by taking the cell count for each confidence level, dividing by the row sum, and cumulating from left to right. A-A' = scenically and mnemonically similar distractor; A-B' = scenically dissimilar but mnemonically similar distractor; A-X = novel distractor.

Response tendency, whereas to do so successfully in the 2AFC case requires a mechanism underlying the improved discriminability.

This is not to say that observers cannot adopt a response bias during a 2AFC task, but to do so in the current task means adopting a tendency to respond to the right or to the left. Given equal presentation probabilities for targets on the left and right, such an outcome would be unlikely, and the adoption of such a strategy would impair, not improve, performance.

Response Confidence

A two-way repeated measures ANOVA with independent variables of distractor condition (A-A' vs. A-B') and response type ("remember" or "familiar") was conducted on confidence for correct responses.1 Results yielded significant main effects for distractor condition, $F(1, 59) = 7.42, MSE = 0.061, p < .01$, and response type, $F(1, 59) = 613.57, MSE = 0.273, p < .001$. An important result was that there was a Distractor condition × Response Type interaction, $F(1, 59) = 6.89, MSE = 0.070, p < .05$. Post hoc tests demonstrated that when participants responded on the basis of feelings of familiarity, responses in the A-B' condition ($M = 2.08$) were more confident than those in the A-A' condition ($M = 1.91$), $t(59) = 3.17, p < .01$. However, confidence did not differ across distractor conditions for remember responses ($t < 1$; see Table I and Figure 2).

Overall, the accuracy and confidence interactions represent two dissociations as a function of response type. Although differences in accuracy between A-A' and A-B' conditions occurred only in remember responses, differences in response confidence were carried solely in familiarity responses. The implications of this result for unidimensional-strength accounts of accuracy-confidence inferences are further examined in the Discussion section.

Dual Criteria for A-A' and A-B' Conditions?

Recently, Donaldson (1996; see also Hirshman & Master, 1997) provided evidence that suggested that previously observed differences in remember–know performance during numerous memory experiments need not be indicative of separate memory systems or states of awareness. Instead, he proposed that many of the differences were more parsimoniously characterized within a simple signal-detection account, assuming that observers adopt two different response criteria for "remember" and "know" judgments. From this perspective, "remember" responses simply reflect conservative responding on the part of participants during yes–no recognition tasks. Again, because criterion placement is independent of accuracy during yes–no procedures, criterion-free estimates of memory accuracy, such as $d'$ or $A'$, should not differ between remember responses and total accuracy for the same condition (see Donaldson, 1996, for further discussion). To examine this, we again looked at the appropriate index of accuracy for forced choice responding: the percentage correct.

Dependent $t$ tests conducted on percentage correct revealed higher accuracy for "remember" responses than overall recognition for A-A' (.88 vs. .70), $t(59) = 12.56$, $p < .001$, A-B' (.74 vs. .66), $t(59) = 6.15$, $p < .001$, and A-X conditions (.94 vs. .81), $t(59) = 10.24$, $p < .001$.

Overall, these data suggest that the basic account of Donaldson (1996) does not hold for the current 2AFC recognition task, quite possibly because the establishment of a conservative response criterion in terms of a difference in target–distractor strength is analogous to an improvement in accuracy in such a task.

---

1 An analogous analysis was not conducted using incorrect "remember" responses. Many participants did not commit any remember errors (other than in the A-B' condition), and those that did made a few at most. Therefore, the averages for each participant were considered unifit for statistical analysis.
Discussion

Our results replicate and extend the findings of Tulving (1981). A more important point is that they suggest that the observed confidence–accuracy inversion observed in this paradigm is not independent of retrieval state and awareness of that state during testing. In cases where participants claimed to remember either the entire scene or elements of the scene, the inclusion of a distractor that was scenically and mnemonically similar (A–A’ condition) yielded higher accuracy than that of a distractor that was scenically dissimilar but similar to a scene previously encountered (A–B’ condition). This was not the case for judgments based solely on feelings of familiarity, for which accuracy did not differ. The reverse pattern was obtained for response confidence. Participants were more confident during the A–B’ condition than the A–A’ condition for responses made on the basis of item familiarity. However, no difference in confidence was observed when participants correctly chose an item on the basis of remembering.

Overall, these results may prove difficult to account for entirely within the signal-detection model proposed by Clark (1997). As discussed earlier, this model assumes item selection is a function of a familiarity difference distribution. When this distribution is assumed to arise from a signal-to-noise ratio, confidence–accuracy inversions are predicted (see Figure 1). What is not predicted is that the confidence and accuracy differences between A–A’ and A–B’ conditions will be manifested for different subjective memory states. Indeed, if difference distributions such as those shown in Figure 1 govern performance in this task, then confidence–accuracy inversions would be expected regardless of the participant’s introspections. This is not to say that the accuracy data during remembering could not be described by such a model; however, the model itself would not shed any light on why this result should only be observed during “remember” responses.

Furthermore, the account may face difficulty simultaneously explaining “remember” performance in A–B’ and A–X conditions. In both cases, participants see scenically dissimilar targets and distractors. If they establish a fixed criterion above which they report remembering, it should be equivalent across conditions. However, given that the overall accuracy is considerably greater for A–X than A–B’, which would be expected because of the novel distractors, one would expect that the proportion of “remember” responses would also be greater in the A–X condition (see Figure 1). However, they are statistically equivalent, and this result cannot occur if they share a common memory criterion. Although one could assert that the conditions entailed different criteria, it is not clear how participants could reliably determine whether they were in an A–B’ or A–X condition. Indeed, examining the ROCs for “familiarity” responses in Figure 2 suggests that participants are using similar criteria in this case, and therefore any proposed criterion differences would have to occur only when responses were accompanied by reports of remembering.

Part of the difficulty may lie in the use of a single-difference distribution to describe remembering in a task with two candidates. Because a difference in strength is a single value, it is not clear how it would reliably map onto the three potential remember outcomes in the current study, namely, (a) neither alternative evokes remembering, (b) both alternatives evoke remembering, or (c) only one alternative evokes remembering. What may be needed is a theory that distinguishes between the two sources of memory judgment (i.e., recollective- and familiarity-based). One set of candidates are dual-process models of recognition that suggest that the two sources can contribute differently to memory performance (e.g., Jacoby, 1991; Mandler, 1980; Yonelinas, 1994). Specifically, the dual-process signal detection of Yonelinas (1994) holds that recognition judgments are the joint product of a continuous, equal variance signal-detection process (accompanied by the subjective feeling of familiarity) and a threshold associative retrieval process (accompanied by a remembering or recollective state; for review, see Yonelinas, Dobbins, Szymanski, Dhalwal, & King, 1996). Furthermore, the model of Jacoby (1991) and colleagues stresses the importance of recollection in controlled or strategic responding. Because targets and distractors are assumed to have equal strength in the A–A’ and A–B’ conditions (Clark, 1997), one would predict equal accuracy in the absence of the contribution of recollection. Both the data analysis and the observed ROCs support this prediction (see Figure 2). This situation is not the case when reports are accompanied by remembering, which is considered to result in a threshold or discrete state of awareness. In this case, participants would be in one of three remembering states on any given trial: (a) They could remember both items, (b) they could remember neither item, or (c) they could remember a single item. The first two states would lead to “familiar” responses because either no information is available, or the information is ambiguous. The last state would contribute solely to “remember” reports.

For example, if we take the A–X remember performance (see Table 1) to accurately reflect the ability to remember previously seen photographs, we can predict the probability of a correct “remember” response in the A–B’ condition. In the case of incorrect A–B’ responses, we assume that participants will only commit such an error if they do not remember the target and yet falsely remember the distractor or if they guess incorrectly on the basis of preexperimental exposures, keyboard errors, and so forth. An important point is that in those cases where they remember both items, they are expected to give a “familiar” response because recollective experience is not useful in the discrimination task. Thus, the probability of an incorrect “remember” response in A–B’ is given by

$$P_{error}(A-B') = P(not A) \times P(B') + g,$$

where $P$ = percentage and $g$ = guessing. Using the data

---

2 This is not to say that the processes underlying remembering for each alternative might not be adequately described by a continuous, signal detection-process (e.g., Donaldson, 1996), merely that the outcome results in three distinguishable cognitive states.
from the A–X condition gives us

$$P_{err}(A-B') = (1 - .46) \times P(B') + .03. \quad (2)$$

Solving for $P(B')$ yields

$$P(B') = .14/.54 + .03/.54 = .20. \quad (3)$$

Therefore, the probability of falsely remembering a distractor is .20. In the case of correct "remember" responding in the A–B’ condition, participants will endorse the correct item if they remember it and do not falsely remember seeing the distractor. Again, assuming independence, this value is .46 × (1 − .20), which is .37 and closely matches the observed result (see Table 1). However, when such an approach is tried for A–A’ “remember” responses, it overestimates the proportion of correct “remember” responses (.39 vs .32). Part of the reason for this failure may lie in the fact that the above calculations rely heavily on the assumption that on any given trial, the probability of remembering the target and of incorrectly remembering the distractor are independent. This situation is unlikely given that the target and distractor are both from the same scene in A–A’ conditions and were selected to be similar in content.

The question remains, why as one goes from the A–B’ to the A–A’ condition does the overall probability of giving a “remember” response drop, yet the accuracy of these responses increase? Similar to Clark (1997), we believe the A–A’ condition results in statistical dependence for memory. That is, because the targets and distractors are drawn from the same scene, the likelihood that a participant will remember one given that they remember the other is substantially increased. This situation would result in two related effects. First, the use of “remember” responses would decline because participants would not report their selection was based on remembering when it is not diagnostic. In those instances where both alternatives are remembered, participants will instead be reduced to relying on feelings of familiarity during selection. Second, accuracy for “remember” reports will improve because many of the most scenically similar pairs will now have been relegated to “familiar” responses. Thus, highly scenically similar distractors of the type that were effective in generating false “remember” reports in A–B’ would be selected against during “remember” reports in A–A’. Subsequently, selection in A–A’ will be made on the basis of information that is more distinctive than in the A–B’ condition, and “remember” accuracy would improve. This result might lead one to predict a concurrent decline in accuracy for “familiar” responses in the A–A’ condition because the more similar pairs have now been relegated to this response category. However, this would only occur if the encoding of stimulus similarity was identical for remembering and familiarity processes, or if such reports relied on identical representations. The long history of dissociations between “remember” and “know” reports (Rajaram & Roediger, 1997) evidence supporting the independence of recollection and familiarity processes (Jacoby, Yonelinas, & Jennings, 1997), and the current set of results weigh against this notion. In addition, research documenting the inability of observers to detect large changes to objects within scenes between viewings may be pertinent.

Simons and Levin (1997) review numerous studies documenting this change blindness that consistently demonstrate that observers are unlikely to notice changes to an object within a scene unless the gist is violated, or the object is likely to have received some degree of prior elaborative encoding (see also Levin & Simons, 1997; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989). Presumably, the ability to detect and describe such changes requires recollective processes, and the difficulty that observers typically have with these tasks suggests that scene recollection operates on relatively sparse or abstract representations. This stands in contrast to traditional accounts of familiarity processes that have often minimized the importance of declarative information and stressed the importance of overlap between the structural or physical properties of stimuli. Although admittedly speculative, future investigation of such representational accounts may yield important information beyond what can be gleaned solely from the dissociation of behavioral response types (see also Eichenbaum, Otto, & Cohen, 1994).

The current levels of selection account stresses the selective use of remembering during responding (e.g., Jacoby, 1991) and suggests that in the A–A’ condition, participants are more selective with regards to the level or distinctiveness of recollected information on which they rely for “remember” reports (see also Rajaram, 1996). Such characterizations are largely absent in traditional remember-know research based on yes–no procedures (e.g., Gardiner & Java, 1990), in part because such designs usually do not place remembering in conflict or opposition. Instead, reports of remembering during yes–no procedures document the presence of a particular memorial state and the efficacy of particular variables in increasing or reducing the occurrence of such a state. The diagnostic value of remembering is typically not in question. During 2AFC responding, however, it is not sufficient to assume the occurrence of remembering during a trial will map directly onto self reports if both alternatives are capable of generating such a state.

In conclusion, the two major outcomes of the current study were the following: (a) “Remember” and “familiar” responses doubly dissociate across A–A’ and A–B’ conditions with familiarity carrying the confidence difference between A–A’ and A–B’ conditions and remembering the accuracy difference; (b) despite lower accuracy, performance in the A–B’ condition was associated with a greater proportion of “remember” responses, with participants often claiming to remember items they in fact did not see. Thus, somewhat paradoxically, the reduced accuracy in the A–B’ condition results from an increase in the number of trials where the subjects are confident that they are remembering one of the alternatives. The current pictorial paradigm is similar to Roediger and McDermott (1995) in which participants often falsely claim to remember high semantic associates of previously encountered word lists. Thus,
participants not only confuse high semantic associates but scenic associates as well.

These results demonstrate that both the tendency to provide “remember” reports and the accuracy of those reports are critically affected by context, namely, the scenic similarity of the distractor. The accuracy of reports based on feelings of familiarity seem to be immune to this effect; however, participants are less confident of their selections when the distractor closely resembles the target. If one assumes that recollection and familiarity processes are independent, such results are interpretable and suggest that the two processes may rely on different or only partially overlapping information. This point in turn suggests that for complex material, such as pictorial scenes, the concept of similarity as a unidimensional or singular property of the stimulus is inappropriate (cf. Tulving, 1981).

The current study offers additional support for the separation of continuous strength–familiarity and discrete recollection processes. However, the simple occurrence of recollection in no way guarantees its task utility, which is critically constrained by the distinctiveness of the retrieved information. One area that remains to be discussed is the degree to which the current effect may be viewed as automatic or volitional. The levels-of-selection account offered in the introduction and drawn primarily from Tulving (1981) suggested that at retrieval participants intentionally engage in deeper or more effortful retrieval strategies when the alternatives are from the same scene. The current dual-process account, however, does not require one to posit such an increase in effort, provided the processes are held to be relatively independent and the construct of distinctiveness is used (e.g., Dobbins, Kroll, Yonelinas, & Liu, 1998). Indeed, given the brief study time of the photographs, it may be unlikely that even large increases in retrieval effort would yield noticeable returns because the required information has simply not been encoded. The previously cited research regarding change detection (e.g., Simons & Levin, 1997) supports this view.

Outside of the laboratory, the difficulty many of us have faced locating our automobile in shopping mall parking lots suggests that scene- or gross-level information is often relied upon by observers until it becomes readily apparent that it is inadequate. At that point, many of us are reduced to either serial search or trial and error. However, this may represent more of a peculiarity of parking lots than a recognition deficiency. Indeed, natural scenes often differ considerably in the fine analysis when approached from different angles or at different times of day, yet it would be unfortunate if this precluded remembering, especially if the location was previously associated with an unfortunate outcome. In short, a large degree of generalizability in scene remembering may typically be a good thing even if people are occasionally led confidently astray.

References


1 The fact that verbal reports during remember–know procedures are necessarily exclusive does not demand any particular relationship between theorized underlying processes. We have adopted an assumption of independence because it is based on previous work and because of its current explanatory value.
Call for Nominations

The Publications and Communications (P&C) Board has opened nominations for the editorships of the Journal of Abnormal Psychology, Journal of Comparative Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition, Journal of Personality and Social Psychology, Attitudes and Social Cognition, Professional Psychology: Research and Practice, Psychological Review, and Psychology, Public Policy, and Law for the years 2001-2006. Milton E. Strauss, PhD; Charles T. Snowdon, PhD; James H. Neely, PhD; Arie W. Kruglanski, PhD; Patrick H. DeLeon, PhD, JD; Robert A. Bjork, PhD; and Bruce D. Sales, JD, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2000 to prepare for issues published in 2001. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

To nominate candidates, prepare a statement of one page or less in support of each candidate. Send nominations to the attention of the appropriate search chair—

- David L. Rosenhan, PhD, for Journal of Abnormal Psychology
- Lauren B. Resnick, PhD, for Journal of Comparative Psychology
- Joe L. Martinez, Jr., PhD, for JEP: Learning, Memory, and Cognition
- Sara B. Kiesler, PhD, for JPSP: Attitudes and Social Cognition
- Judith P. Worell, PhD, for Professional Psychology: Research and Practice
- Lyle E. Bourne, Jr., PhD, for Psychological Review
- Lucia A. Gilbert, PhD, for Psychology, Public Policy, and Law

—to the following address:

c/o Karen Sellman, P&C Board Search Liaison
Room 222
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242

The first review of nominations will begin December 7, 1998.