The face typicality–recognizability relationship: Encoding or retrieval locus?

KENNETH A. DEFFENBACHER and JOHN JOHANSON
University of Nebraska, Omaha, Nebraska

THOMAS VETTER
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

and

ALICE J. O'TOOLE
University of Texas, Dallas, Texas

Using a crossover recognition memory testing paradigm, we tested whether the effects on face recognition of the memorability component of face typicality (Vokey & Read, 1992, 1995) are due primarily to the encoding process occurring during study or to the retrieval process occurring at test. At study, faces were either veridical in form or at moderate (Experiment 1) or extreme (Experiment 2) levels of caricature. The variable of degree of facial caricature at study was crossed with the degree of caricature at test. The primary contribution of increased memorability to increased hit rate was through increased distinctiveness at study. Increased distinctiveness at test contributed to substantial reductions in the false alarm rate, too. Signal detection analyses confirmed that the mirror effects obtained were primarily stimulus/memory-based, rather than decision-based. Contrary to the conclusion of Vokey and Read (1992), we found that increments in face memorability produced increments in face recognition that were due at least as much to enhanced encoding of studied faces as they were to increased rejection of distractor faces.

The rated typicality of faces has perhaps been found to be the most reliable predictor of how well observers come to recognize them (e.g., Bartlett, Hurry, & Thorley, 1984; Light, Kayra-Stuart, & Hollander, 1979), with typical faces being less well recognized than atypical or unusual ones. This effect is due both to a higher hit rate and to a lower false alarm rate for atypical or distinctive faces, as has been documented in Shapiro and Penrod's (1986) meta-analysis of face recognition memory studies ($d = .76$ for hits, $d = .78$ for false alarms). The cited effect sizes are moderate to large by Cohen's (1977) criteria.

As Vokey and Read (1992, p. 302) have pointed out, the concept of face typicality is clear enough as a theoretical construct, but its instantiation varies considerably across studies. In some cases, typicality of a face is defined with explicit reference to a prototype. Light et al. (1979, Experiment 1) asked their raters to compare how similar each face was to their "idea of the typical high school male

senior." Vokey and Read (1992, Experiment 1) asked their observers to rate the typicality of each face with reference to how sure they were that it was "average." Other investigators have required their observers to rate the difficulty of picking a given face out of a crowd (O'Toole, Deffenbacher, Valentin, & Abdi, 1994; Valentine & Bruce, 1986). Still others (Ellis, Shepherd, Gibling, & Shepherd, 1988) have equated face typicality/distinctiveness with the degree of objective memorability of a set of faces. We agree with Vokey and Read (1992) that these different measures of face typicality may not necessarily capture equally well the underlying theoretical construct. However, as we shall note shortly, there is evidence that the various instantiations may be generally compatible.

Vokey and Read (1992) undertook a structural analysis of human observer ratings of face typicality, familiarity, memorability, likableness, and attractiveness, using principal components analysis, and examined how different components of this structure related to various measures of face recognition. Interestingly, they found that typicality is a more complex concept than had previously been supposed. Vokey and Read (1992) showed rather convincingly that typicality is composed of two orthogonal components. One component, directly related to face memorability (how easy observers thought a face would be to remember), was found to be a significant predictor of both hits (direct relation) and false alarms (inverse relation). Henceforth, for brevity, we will refer to this com-

ponent as the memorability component of typicality. The other component, directly related to rated familiarity (the extent to which observers thought a face might have been seen around campus), attractiveness, and likability was labeled *context-free* familiarity by Vokey and Read (1992). This component significantly predicted false alarms to faces on a recognition memory test, but not hits.

O'Toole et al. (1994) also examined structural aspects of face typicality, for both same-race and other-race observers. They replicated Vokey and Read's (1992) findings for Caucasian observers viewing Caucasian faces. However, face typicality of Japanese faces as viewed by Caucasian observers was related only to memorability. This indicates that observers may judge the typicality of other-race faces with different criteria than they do in judging own-race typicality. For present purposes, however, the replication of Vokey and Read (1992) by O'Toole et al. (1994) for same-race faces indicates that the different operational definitions of typicality used in the two studies were equivalent.

The general, or context-free, familiarity of faces has been hypothesized to be related to the similarity space of faces, one based on a lifetime of experience with similar faces (Light et al., 1979, p. 225). It has the effect of increasing the likelihood of false alarm responses to distractor faces in recognition memory tests. As Mandler (1980) and others have theorized, the experience of familiarity does not directly index source. Hence, unfamiliar faces high in structurally induced familiarity (general, or context free) are difficult to discriminate from faces with lower structurally induced familiarity but with which there has been at least one specific episode of encounter.

By contrast to the context-free component of typicality, the memorability component is thought to be "bound" to a context or to an episode of encounter. It is furthermore thought to relate to the ease of encoding atypical versus typical faces. From the strength models of recognition, distinctive faces emerge from an episode of encounter with a greater increment of memory strength than do more typical faces. For example, Light et al. (1979) found that the subsequent recognition of atypical faces benefited more from elaborative encoding tasks than did recognition of more typical faces. Bartlett et al. (1984) also found that specific prior exposure to atypical faces enhanced their recognition discriminability more than it did for typical faces.

A face with high memorability is thought to elicit fewer false alarms, because in the absence of retrieval of a good match to the distractor face, observers are more likely to reject it. They reject it with greater confidence, owing to its high assessed memorability: Lack of retrieval for what should have been an easily retrieved face is good evidence that it should not be accepted as a target (Brown, Lewis, & Monk, 1977; Vokey & Read, 1992).

There is also some evidence that the memorability component of face typicality can be traced to the presence of small, local distinctive features in some of the faces observers consider atypical (O'Toole et al., 1994).

As was noted, O'Toole et al. (1994) found that typicality related only to the memorability component for other-race faces, suggesting that structural or context-free familiarity applies only to same-race faces. This is because our psychological face space is derived from the statistical structure of our experience with faces, a structure usually strongly biased to faces of a single race. Note that this explanation of memorability is not inconsistent with those offered previously. Rather, it localizes the benefits of face distinctiveness during an encounter at the level of the information in the face stimulus.

Having explored the basis for the effects of typicality on face recognition, it is of interest to try to determine where these effects operate—during encoding processing, retrieval processing, or both. Even were typicality to produce its effects at both stages of processing, it could well be that the effects might not be symmetrical. Vokey and Read (1992), for instance, have concluded that the principal benefit of distinctiveness/memorability to improved face recognition is not its ability either to enhance the encoding process at study or to provide more distinctive retrieval cues at test. Rather, the greater contribution of increased memorability is to reduce the false alarm rate. Presumably, highly distinctive/memorable *new* faces presented at memory test are less likely to be highly similar to any faces in memory. They are therefore less likely to be a good match to any candidate face retrieved and are thus rejected as *old* with greater confidence, owing to their obvious high memorability.

Although Vokey and Read's (1992) data were reasonably consistent with their conclusion that face memorability's principal contribution to improved face recognition is to the rejection of *new* faces as *old*, we are not convinced that memorability's effect is quite so asymmetrical in favor of more efficient retrieval processing of *new* faces at test. As Vokey and Read (1992, Experiment 3) acknowledged, at least part of their failure to find much effect of memorability on the hit rate may have been due to their use of memorability component scores of *test* faces, rather than of *study* faces. We suspect that when one directly manipulates the degree of face distinctiveness/memorability at both study and test, one will find stronger evidence of its contribution to improvements in the hit rate, whether by enhanced encoding distinctiveness or by enhanced retrieval cues at test.

We can test this conjecture by use of a crossover recognition memory test (see, e.g., Fiorentini, Maffei, & Sandini, 1983; O'Toole, Millward, & Anderson, 1988). The simplest and most elegant experiment would be to present both veridical faces and faces of artificially enhanced memorability (distinctiveness) at study and then to cross this within-subjects variable with another within-subjects variable, veridical faces or faces of enhanced memorability at test. Suppose there were 40 study faces. Ideally, one would have an observer study 20 unenhanced or veridical faces and 20 faces of enhanced memorability. At test, 10 of the veridical faces would be presented in the same unenhanced state, but the other 10 of them would be

presented in a format of enhanced distinctiveness or memorability. Similarly, 10 of the originally enhanced targets would be presented in the same format at test, and the other 10 would appear in their unenhanced, or veridical, form.

Consider what the results of a two-factor analysis of variance (ANOVA) of accuracy data (e.g., A' scores) from such an experiment could tell us about the locus of the effect of face memorability/distinctiveness.[1] Were the effect only to occur during the process of encoding (a process operating at study), the ANOVA would reveal just one significant effect, a main effect of distinctiveness at study, so that recognition for faces studied in their more distinctive form would be better than that for faces studied veridically. A main effect of testing condition only would be found were the effect of facial distinctiveness to occur just during retrieval processing. Here, regardless of face distinctiveness at study, it would be advantageous for recognition to have its more distinctive form present at test. It is important to note that although these two possible results would be clear indication of the dominance of one or the other locus of the effects of the memorability/distinctiveness component of typicality, such outcomes may be a bit simplistic.

For instance, consider the case in which the only significant result from the two-factor ANOVA would be a complete crossover interaction. Here, the source of the interaction would be the encoding specificity phenomenon (Tulving & Thomson, 1973). Accuracy of recognition memory would be better for the two conditions in which the study and the test faces were identical, regardless of their enhancement status.

In order to conduct an experiment such as the one being proposed, we need a means of manipulating the degree of distinctiveness/memorability of faces. Recently, we have demonstrated the feasibility of doing so by applying a standard computer caricature algorithm (see Benson & Perrett, 1991; Brennan, 1985; Rhodes, Brennan, & Carey, 1987) to three-dimensional (3-D) representations of human faces taken from laser scans (Deffenbacher, Vetter, Johanson, & O'Toole, 1998; O'Toole, Vetter, Volz, & Salter, 1997). The algorithm operates by changing the locations of individual faces relative to an average face in a multidimensional face space and by "redrawing" the caricatured face with the altered feature values. With this algorithm, we first represent individual faces as points or vectors in a multidimensional space on the basis of their 3-D representations. Each face can be characterized by a distance and direction from the average face. The distance of a face from the average face is an empirically supported measure of its distinctiveness and memorability (Deffenbacher et al., 1998),[2] whereas the direction of the face vector captures its identity, via vector values on the feature dimensions defined by the axes of the face space. Finally, caricatures are created by increasing the length of a face's vector, thereby altering its distinctiveness/memorability and distance from the average face, while leaving its direction (identity) intact. Vector lengths are measured as

Mahalanobis distances (Duda & Hart, 1973) or, in more familiar terms, z score units.[3]

In the present study, we conducted two experiments to test the hypothesis that the memorability component of face typicality exerts strong effects on face recognizability during the process of encoding, as well as during retrieval. In the first experiment, we test this notion with a crossover recognition memory design, using faces both in their veridical form and in a form in which their distinctiveness/memorability has been computer enhanced by an amount we would describe as *moderate*. In the second experiment, we repeat the test with faces having both a veridical form and a form with a memorability enhancement we would describe as *extreme*.

EXPERIMENT 1

Here, we tested whether the memorability component of face typicality has a stronger effect on encoding processes than the results of Vokey and Read (1992) would lead one to suspect. We presented a set of study faces, half veridical and half with their memorability and distinctiveness artificially enhanced to a moderate degree. Degree of face memorability/distinctiveness at study was then crossed with the variable of degree of face memorability/distinctiveness at test, yielding four transfer conditions. We predicted that the most likely pattern of results would include a main effect for study condition but no main effect for test condition, owing to an associated crossover interaction, involving small but symmetrical encoding specificity effects. Faces studied in their enhanced form would be recognized with greater accuracy than faces studied in their unenhanced form. However, at test there would be a modest accuracy advantage for faces both studied and tested in their enhanced form, as compared with faces studied in their enhanced form but tested in their veridical form. A similar modest advantage would accrue to faces both studied and tested in their veridical form, relative to faces studied in their veridical form but tested in their enhanced form.

Method

Participants. The participants were 25 undergraduates (20 female, 5 male) enrolled in psychology courses at the University of Nebraska at Omaha. They received course extra credit for their participation. They were exposed to all four conditions of a 2 (study condition: veridical faces or moderately caricatured faces) × 2 (test condition: veridical faces or moderately caricatured ones) within-subjects factorial design.

Stimuli. The stimuli were laser scans (Cyberware TM) of 64 heads of young Caucasian adults, 32 of each sex. These were randomly selected from a database of 100 faces (mean age = 26.9 years, standard deviation = 4.7 years). Individuals were scanned wearing bathing caps, which were later removed digitally. Further preprocessing of the heads involved making a vertical cut behind the ears and a horizontal cut to remove the shoulders. The laser scan data consisted of a representation of the 3-D coordinates of a 512×512 sample of head surface points. A more complete description of the laser data can be found in O'Toole et al. (1997).

Figure 1. Head of a 26-year-old male in its veridical form (left), at a distance from the average head that corresponds to a moderate degree of caricature (center), and at a distance corresponding to an extreme caricature (right).

The 3-D face data were rendered as images viewed from 30° left of full-face view. Figure 1 displays the face of a 26-year-old male in its veridical form and in two levels of increasing caricature. In the present study, faces were presented either in their veridical form (average Mahalanobis distance from the average face of 10.0) or in a form wherein the distance from the average face had been increased by the caricature algorithm to an average distance of 13.5, producing a moderate degree of caricature. As was noted previously, Deffenbacher et al. (1998) have shown that a distance of 13.5 from the average face produces significantly higher ratings of distinctiveness and a significantly greater degree of memorability than does an average distance of 10.0.

Apparatus. Individual testing of each participant was accomplished by having all the experimental events controlled by a Macintosh computer programmed with PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993).

Procedure. In pilot testing for a previous study (Deffenbacher et al., 1998), we had noted that a caricature effect was not obtainable in recognition memory for previously unfamiliar faces unless there was a minimum of about 60 sec of prior exposure to each face. Consequently, in the present study we had the participants engage in a rating task wherein they made four ratings of each face, using 5-point, Likert-type scales. During each of the four rating tasks, the 32 study faces (eight males and eight females at an average distance of 10.0 from the average face, plus eight males and eight females at a distance of 13.5) were presented in a unique random order for each participant at a 15 sec per image rate, with a 1-sec interstimulus interval (ISI). The order of the rating tasks was to rate all study faces first for attractiveness, then for distinctiveness, then for honesty, and finally for likeability. The participants were reminded that they should study each face carefully, given that their memory would be tested in the second phase of the experiment.[4]

A recognition memory test occurred immediately after the completion of the fourth rating task. Each participant viewed 64 faces in a unique random order, all of the 32 study faces (*old*) and 32 *new* distractor faces. Of the 32 *old* faces, 16 remained in the same format as that at study (8 at a distance of 10.0 from the average face, 8 at a distance of 13.5). The other 16, however, were presented in the alternate format at test. Eight were studied in the 10.0 format and were tested in the more distinctive/memorable 13.5 format. Eight others were studied in the 13.5 format but were tested in the veridical, 10.0 format. Of the 32 *new* faces, 16 were presented in each of the two formats, there being equal numbers of male and female faces in each format.

The participants were instructed that their task was to determine whether each face was *old* or *new*. Since some faces were shown in a different format at test, the participants were informed that some

of the *old* faces had been electronically altered to look slightly older or younger but were, in fact, still the same faces. Each participant was to press "1" on the computer keyboard if he/she believed that a given face had been seen before and to press "2" if not. The faces were presented at a 5-sec rate, and the participant was allowed to respond anytime during or after this interval. A 1-sec ISI followed presentation of each face or each response, whichever occurred last.

Results and Discussion

Analyses of affirmative responses. The effects of study and test conditions on mean proportion of *yes* responses and corresponding standard error are illustrated in Table 1. Because the two false alarm rate conditions in the top row were each based on twice as many trials ($n = 16$) as were each of the four hit rate conditions in the bottom two rows, and because the primary test of our predictions depended on an analysis of the four hit rate conditions, we conducted separate analyses of false alarm and hit rate conditions.

Given that the faces did not serve in all conditions of the experiment, parallel analyses of hit and false alarm rates were conducted, first with observers as the random variable and then with faces as the random variable. In every instance, the results of the two analyses were identical in terms of statistical significance. Hence, we have chosen to report only the results of the former analyses.

As Vokey and Read (1992, 1995) would have predicted, a comparison of the false alarm rate conditions revealed that more distinctive (13.5) *new* faces elicited a significantly lower false alarm rate than did *new* faces in their unaltered format (10.0) [$F(1,24) = 24.00, p < .001$]. The effect size associated with this 47% decrease in false

Table 1
Mean Proportion of *Yes* Responses and Associated Standard Errors as a Function of Study and Test Condition in Experiment 1

Condition	Tested Veridical		Tested Caricature	
	M	*SE*	*M*	*SE*
Not studied	.268	.030	.142	.026
Studied veridical	.835	.020	.630	.033
Studied caricature	.900	.018	.915	.022

Table 2
Mean A' Scores and Associated Standard Errors as
a Function of Study and Test Condition in Experiment 1

Condition	Tested Veridical		Tested Caricature	
	M	SE	M	SE
Studied veridical	.859	.016	.787	.025
Studied caricature	.910	.011	.933	.013

alarm rate was moderately large ($r_{pb}^2 = .50$).[5] Clearly, this effect of memorability/distinctiveness on the false alarm rate is one that could have occurred only during retrieval processing associated with recognition memory testing. Again, *new* 13.5 faces should have been easier to reject as *old* than *new* 10.0 faces, given the higher assessed memorability of the former faces. Lack of retrieval of a good match to a 13.5 distractor face, one that should have been easily retrieved had it been encountered previously, would have been a good reason to reject it as a target face.

A 2×2 within-subjects ANOVA was then conducted on the hit rate scores. There was a sizable main effect for study condition [$F(1,24) = 81.10, p < .001, r_{pb}^2 = .77$]. Studying faces in their enhanced form produced an average hit rate of .908, clearly superior to the average hit rate of .732 obtained when faces were studied in their unaltered, veridical form. Both simple effects for study condition were significant in this instance. When faces were tested in their veridical form (10.0), there was an advantage to having studied them in a moderate caricature (13.5) [$F(1,24) = 9.99, p < .01$]. Thus, veridical faces were better recognized when they had been studied in a modestly more distinctive/memorable format than when they had been studied in their veridical version.[6] There was an even stronger effect of study condition when faces were tested in their more distinctive format (13.5) [$F(1,24) = 84.21, p < .001$].

Contrary to prediction, the main effect for test condition was significant [$F(1,24) = 18.18, p < .001, r_{pb}^2 = .43$]. Mean hit rate for faces tested in their enhanced form was actually less than that for faces tested in their veridical form (.772 vs. .868, respectively). This effect is in the direction opposite to that predicted for a retrieval-processing locus for memorability's effect on hits in a face recognition task. The effect is due to the obtained interaction [$F(1,24) = 38.27, p < .001$] not demonstrating the predicted symmetrical effects of encoding specificity. There was a large decrement owing to a lack of encoding specificity for 10/13.5 faces, but none whatsoever for 13.5/10.0 faces. The extra distinctiveness of the 10/13.5 faces at retrieval provided no shield from their associated lack of encoding specificity, whereas the extra distinctiveness at encoding of the 13.5/10.0 faces did so.

Hence, the *primary* locus of the effect on face recognition hits for the distinctiveness/memorability component of face typicality is during the encoding process. The effect size for the main effect for the study condition obtained in Experiment 1 was considerably larger than that for the main effect for the test condition. In fact, the main

effect for the test condition was in the direction opposite to that predicted for a retrieval-processing locus for memorability. Both simple effects for the study condition were substantial, too—an effect size of .78 in the case of faces tested in their more distinctive format and of .29 in the case of faces tested in their veridical format. A graph-oriented way to see that the primary locus of the effect of distinctiveness/memorability on hits was during encoding would be to note that an average of the two curves for the study condition simple effects would be a straight line of moderate positive slope, visual confirmation that the locus of the effect was entirely during the process of encoding.

Thus, the results of Experiment 1 strongly support Vokey and Read's (1992, 1995) hypothesis that the memorability component of face typicality achieves its efficiencies through both increasing hits and decreasing false alarms. It would also appear that moderate increases in distinctiveness/memorability have their effect on *both* encoding and retrieval processing attendant to a face recognition task. Arguably, the effect of memorability on hits during encoding is more pronounced than its effect on false alarms during retrieval, contrary to what Vokey and Read (1992) found. After all, the effect size of memorability on the hit rate at study was noticeably larger (.77) than its effect on the false alarm rate at test (.50). However, this conclusion must remain a tentative one, until we can demonstrate that memorability had independent effects on hit and false alarm rates. For this demonstration, we now turn to analyses of signal detection measures of face recognition, analyses that should allow us to separate stimulus- or memory-based effects of our memorability manipulation from decision- or criterion-based effects.

Signal detection analyses. First, we report the results of a 2×2 ANOVA of A' scores computed from hits and false alarm rates and yielding here a nonparametric index of the discriminability of *old* and *new* faces that is independent of changes in response criterion (Grier, 1971). This index ranges from .50 (chance discrimination) to 1 (perfect discrimination). Mean A' scores and associated standard errors are reported, for each combination of study and test conditions, in Table 2. We should point out that in calculating A', we compared hit rates for the 10.0/10.0 and 13.5/13.5 target faces with false alarm rates for *new* 10.0 and *new* 13.5 faces, respectively. For target faces in the two crossover conditions, 10.0/13.5 and 13.5/10.0, we felt that it was most appropriate to compare their hit rates with the pooled false alarm rate for *new* 10.0 and 13.5 faces. Consider a target face in the 10.0/13.5 condition. The 13.5 version of the face presented at test would certainly warrant comparison with the false alarm rate for 13.5 distractor faces. On the other hand, the observer's memory representation of the face in question would be at the unenhanced (10.0) level of distinctiveness. This would warrant comparison of the hit rate with the false alarm rate for 10.0 distractor faces. Thus, the decision to use the pooled false alarm rate in these two conditions seemed the most reasonable decision to make.

Table 3
Mean B'' Scores and Associated Standard Errors
as a Function of Study and Test Conditions in Experiment 1

Condition	Tested Veridical		Tested Caricature	
	M	SE	M	SE
Studied veridical	−.155	.082	.165	.061
Studied caricature	−.341	.092	−.309	.096

The results of this analysis confirm almost exactly the results of the hit rate analysis. Again, there was a sizable main effect for study condition [$F(1,24) = 54.87, p < .001, r_{pb}^2 = .70$], with faces studied in their enhanced form producing an average A' of .922, clearly superior to the average of .823 obtained when faces were studied in their veridical form. Both simple effects for the study condition were likewise significant again, too ($p < .001$ in both instances); effect sizes were comparable with those obtained in the hit rate analysis. As before, there was a smaller main effect for testing condition [$F(1,24) = 6.08, p < .05, r_{pb}^2 = .20$]. The mean A' score for faces tested in their veridical form was actually superior to that obtained for faces tested in their caricatured form (.884 vs. .860), because the same sort of asymmetrical interaction as that in the hit rate analysis was found [$F(1,24) = 25.72, p < .001$].

Next, we report briefly the results of a 2 × 2 ANOVA of B'' scores, a nonparametric measure of response criterion (Grier, 1971). This index ranges from −1 extreme liberal criterion) to +1 (extreme conservative criterion), with a score of 0 representing no response bias. Mean B'' scores and their associated standard errors are reported in Table 3. Both of the main effects and the interaction were statistically reliable [$F(1,24) = 22.59, p < .001$, for the main effect of study condition; $F(1,24) = 5.32, p < .05$, for the main effect of test condition; and $F(1,24) = 7.49, p < .05$, for the interaction]. In studying faces in their enhanced (13.5) form, the observers adopted a more liberal response criterion than when studying faces in their veridical form. The interaction and the main effect of test were due primarily to the more conservative criterion adopted by observers in the 10.0/13.5 crossover condition, the condition of poorest performance in the hit rate and A' analyses. Perhaps not surprisingly, the observers became more conservative in responding *yes* to 10.0/13.5 target faces at test, given the difficulties encountered in retrieving a less distinctively encoded version of the face and trying to match it to the more distinctive version being presented.

Even though the observers adopted a somewhat more liberal criterion when studying faces of enhanced distinctiveness than when studying veridical faces, this cannot account for the significant effect on hit rate of the distinctiveness/memorability manipulation. That is, simply adopting a more liberal criterion with no effect of the stimulus manipulation on memory for the faces would have increased both the hit and the false alarm rates commensurately, yielding no net effect of study condition in

an analysis of A' scores. Nevertheless, the pronounced positive effects produced both on hit rates and A' scores by studying faces in their enhanced (13.5) form demonstrate that our stimulus manipulation yielded independent effects on hits (increase) and false alarms (decrease). Criterion-based effects associated with our stimulus manipulations at study and test were likewise independent of memory-based effects.

EXPERIMENT 2

We next sought to determine whether the effects observed in Experiment 1 would be accentuated if the crossover recognition memory design used faces, not of moderate caricature, but of a more extreme caricature. Accordingly, we repeated the procedure of Experiment 1, using veridical faces (10.0) and faces at an even greater average Mahalanobis distance (17.0) from the average face in the multidimensional space based on their 3-D representations.

Method

Participants. Again, the participants were 25 undergraduates (17 female, 8 male) enrolled in psychology courses at the University of Nebraska at Omaha. As did the Experiment 1 participants, they received course extra credit for their participation. They were exposed to all four conditions of a 2 (study condition: 10.0 faces or 17.0 faces) × 2 (test condition: 10.0 faces or 17.0 faces) within-subjects factorial design.

Stimuli. The stimuli were the same 64 faces as those used in Experiment 1. The only difference in Experiment 2 was that all the manipulations with 13.5 faces were now manipulations with 17.0 faces, highly distinctive/memorable versions of a face (Deffenbacher et al., 1998).

Procedure. The apparatus and procedure were identical to those in Experiment 1.

Results and Discussion

Analyses of affirmative responses. The mean proportions of *yes* responses and the corresponding standard errors are displayed in Table 4 as a function of study and test conditions. As in Experiment 1, separate analyses were conducted on false alarm rates (top row) and hit rates (bottom two rows).

Not surprisingly, given Vokey and Read's (1992, 1995) results and theorizing and the results of Experiment 1, highly distinctive (17.0) *new* faces produced a very sizable reduction in false alarm rate, as compared with that elicited by the *new* faces presented in their unaltered, veridical format (10.0) [$F(1,24) = 98.67, p < .001, r_{pb}^2 = $

Table 4
Mean Proportion of *Yes* Responses and
Associated Standard Errors as a Function
of Study and Test Conditions in Experiment 2

Condition	Tested Veridical		Tested Caricature	
	M	SE	M	S E
Not studied	.402	.034	.120	.017
Studied veridical	.860	.035	.395	.048
Studied caricature	.855	.031	.970	.011

Table 5
Mean *A′* Scores and Associated Standard Errors as
a Function of Study and Test Condition in Experiment 2

Condition	Tested Veridical		Tested Caricature	
	M	*SE*	*M*	*SE*
Studied veridical	.819	.022	.597	.043
Studied caricature	.873	.015	.960	.006

.80]. Again, this effect could have occurred only during retrieval processing at test. It should be noted, however, that in Experiment 2 most of the accentuated reduction in false alarm rates (as compared with Experiment 1) by the more distinctive distractor faces was due to an interesting context effect. The false alarm rate to the 17.0 faces in this experiment (.12) was only slightly lower than the false alarm rate to 13.5 faces in Experiment 1 (.14). The major difference in false alarm rates between the two experiments, however, was the *increased* false alarm rate to 10.0 faces in the present experiment [.40 vs. the .27 rate in Experiment 1; $t(48) = -2.98, p < .01$, two-tailed]. Apparently, in the context of the extreme caricature for 50% of the target and distractor faces in Experiment 2, unaltered *new* faces (10.0) were perceived as even more familiar than they were in the stimulus context of Experiment 1, in which the same 50% of the target and distractor faces were in a more moderate state of caricature (13.5). This result serves to document the fact that manipulations of face typicality may be sensitive to list context effects, a phenomenon noted recently by Hosie and Milne (1996).

The same ANOVA of hit rate scores was conducted as that done in Experiment 1. There was again a sizable main effect for study condition [$F(1,24) = 70.82, p < .001$, $r_{pb}^2 = .75$]. Studying faces in their enhanced form (17.0) produced a mean hit rate of .912, much superior to the mean hit rate of .628 produced by studying faces in their unaltered form (10.0). Here, unlike Experiment 1, however, only one simple effect for the study condition was significant. When faces were tested in their more distinctive format (17.0), there was a decided advantage to having studied them in that same distinctive format [$F(1,24) = 138.00, p < .001$].

The main effect for the test condition was significant, again, too [$F(1,24) = 24.00, p < .001$]. As in Experiment 1, the mean hit rate for faces tested in their enhanced form was actually less than that for faces tested in their veridical form (.682 vs. .858). The reason for this particular result's being in the direction opposite to that predicted for a retrieval-processing locus for memorability's effect on hits was the same as that for Experiment 1. There was here, too, a significant interaction [$F(1,24) = 106.29, p < .001$] that did not display symmetrical effects of lack of encoding specificity. The lack of encoding specificity for the 10.0/17.0 faces produced a much more serious decrement of the hit rate than did the lack of encoding specificity for the 17.0/10.0 faces. Apparently, the highly distinctive encoding of the 17.0/10.0 faces shielded them somewhat from the effect of a lack of en-

coding specificity, whereas for a face encoded in its unaltered form (10.0), presenting a highly distinctive version at test was of little help.

Although the results in Experiment 2 were perhaps a bit less clear cut than those for Experiment 1, there was still little evidence of a retrieval-processing locus for the action of memorability/distinctiveness on the hit rate. If there had been, the hit rate for the 10.0/17.0 faces should have been much greater than it was. There was, nevertheless, a retrieval locus for face memorability's effect on face recognition, as indexed by its significant lowering of the false alarm rate. As contrasted with Experiment 1, however, the effect of increased distinctiveness/memorability on the hit rate (presumably occurring during the process of encoding) was not more pronounced than its effect on the false alarm rate (during retrieval). As a matter of fact, the effect size in the former instance (.75) was a bit smaller than that in the latter instance (.80). Again, however, before making much of this apparent difference in effect sizes, we need to confirm the independence of enhanced memorability's effects on the hit and the false alarm rates.

Signal detection analyses. *A′* scores were computed as in Experiment 1, and the means and standard errors are reported in Table 5. A 2 × 2 ANOVA of these scores indicated that, with one exception, the exact same set of results was obtained as that for the hit rate analysis. As with the hit rate analysis, a sizable main effect of study condition was obtained [$F(1,24) = 63.52, p < .001$, $r_{pb}^2 = .73$]. Faces studied in their highly enhanced (17.0) form produced a mean *A′* of .916, definitely superior to the mean score of .708 produced when studying faces in their unenhanced (10.0) form. However, although the hit rate analysis only showed one significant simple effect for the study condition, both simple effects were significant here. The hit rates for the 10.0/10.0 and 17.0/10.0 conditions were virtually identical. However, the false alarm rate was lower enough in the latter condition that a significant difference was obtained between these two conditions in the *A′* analysis. Otherwise, there was, again, a relatively small main effect for the test condition [$F(1,24) = 8.96, p < .01, r_{pb}^2 = .27$]. As was the case in the hit rate analysis in this experiment and in comparable analyses in Experiment 1, performance with faces tested in their veridical form was actually better than that with faces tested in their distinctive form, resulting in mean *A′* scores of .846 versus .778. The same sort of asymmetrical crossover interaction was obtained here, too [$F(1,24) = 55.18, p < .001$].

We now report results of a 2 × 2 ANOVA of *B″* scores, in order to check for the presence of criterion shifts attendant to our stimulus manipulations at study and test. Mean *B″* scores and their standard errors are presented in Table 6. As in Experiment 1, both of the main effects and the interaction were statistically reliable [$F(1,24) = 5.93, p < .05$, for the main effect of study; $F(1,24) = 5.05, p < .05$, for the main effect of test; and $F(1,24) = 15.00, p < .01$, for the interaction]. All three

Table 6
Mean B'' Scores and Associated Standard Errors as
a Function of Study and Test Condition in Experiment 2

Condition	Tested Veridical		Tested Caricature	
	M	SE	M	SE
Studied veridical	−.470	.086	−.023	.087
Studied caricature	−.362	.105	−.492	.093

of these effects were due to the fact that the criterion adopted by the observers in the 10.0/17.0 crossover condition was significantly more conservative than any of the response criteria adopted in the other three conditions. Again, this particular crossover condition was associated with the poorest performance in terms of hit rate and A'. It should be noted, of course, that the 10.0/17.0 criterion was conservative only relative to the moderately liberal criteria associated with the other three experimental conditions; this criterion was neutral or unbiased in an absolute sense.

As was the case in Experiment 1, response criterion differences cannot account for the significant effects on hit rate of the distinctiveness/memorability manipulation, effects replicated in the A' analysis. Again, studying faces in their enhanced form resulted in a more liberal response criterion that did studying faces in their veridical form. Indeed, this decision-based difference could account for the obtained greater hit rate for faces studied in their highly distinctive form, in the absence of any memory-based effect of the distinctiveness/memorability manipulation. However, as the separate hit and false alarm analyses showed, the stimulus manipulations acted separately both to increase the hit rate as a result of encoding processing and to decrease the false alarm rate during retrieval processing. This demonstration was confirmed by an analysis of A' scores. Thus, our stimulus manipulations elicited both memory/stimulus-based effects on face recognition and decision-based ones. The latter were independent of the former, however.

GENERAL DISCUSSION

Across both experiments, we found compelling evidence in support of Vokey and Read's (1992, 1995) hypothesis that the memorability component of face typicality acts both in a direct relation with the hit rate and in an inverse relation with the false alarm rate. Thus, for face recognition memory tasks, increases in the memorability component of face typicality act both to increase hits and to reduce false alarms.

What can we conclude concerning the relative contributions of memorability/distinctiveness to the encoding and to the retrieval processing accompanying face recognition tasks? It is clear that the primary contribution to increases in the hit rate comes from increases in encoding distinctiveness. Any contribution of increases in distinctiveness at retrieval to increases in the hit rate would certainly have elevated the hit rate in the 10.0/13.5 and

10.0/17.0 transfer conditions so that there would have been much less of an encoding specificity deficit than was shown. It is likewise clear that increases in distinctiveness during retrieval processing contribute to substantial reductions in the false alarm rate. As contrasted with the results of Vokey and Read (1992), however, we have shown a relatively greater contribution of distinctiveness/memorability to the effectiveness of encoding processing. After all, the effect size for face memorability's contribution to encoding distinctiveness (increase in hits) was fully 54% larger in Experiment 1 than was its contribution to more efficient processing at retrieval (decrease in false alarms). The squared point–biserial correlation coefficients were .77 and .50, respectively. This relative advantage was reversed under the conditions of Experiment 2, where the greater absolute disparity in distinctiveness/memorability between faces in their unaltered and enhanced forms led to the previously cited list context effect; here the unaltered *new* faces appeared to be perceived as more generally familiar than they were in Experiment 1, thereby accentuating the difference in false alarm rates between unaltered and enhanced *new* faces. Nevertheless, the effect size for enhanced encoding distinctiveness in this instance was only 6% smaller than that for the enhanced retrieval processing promoted by the presence of highly distinctive *new* faces, the squared point–biserial coefficients being .75 and .80, respectively.

These effects, obtained for enhanced memorability/distinctiveness on both hit rates and false alarm rates, likewise indicate the presence of a *mirror effect* (Glanzer & Adams, 1985). The mirror effect refers to the robust phenomenon in recognition memory whereby classes of stimuli easier to recognize when *old* are also easier to categorize as *new* when they have not been previously encountered. Many previous investigators of face recognition memory have reported what amounts to a mirror effect, although it is rarely acknowledged as such, Vokey and Read (1992, 1995) being the first to do so. These previous investigators have reported results demonstrating the usual typicality effect for face recognition: Typical faces have been less well recognized than atypical or distinctive ones, with this effect being due to both a higher hit rate and a lower false alarm rate for atypical or distinctive faces. Indeed, these results have been nicely summarized by Shapiro and Penrod's (1986) meta-analysis of face recognition memory studies, cited in the introduction to this paper.

Even though we did not configure these two experiments so as to assess directly the validity of models that have been proposed to account for the mirror effect, we are nevertheless able to provide some suggestions as to why our distinctiveness/memorability manipulation gave rise to mirror effects in both the experiments. At least for verbal stimuli, Stretch and Wixted (1998) have shown that some mirror effects occur as a result of strength manipulations (study time, for instance); they suggest that these mirror effects are due to shifts in the decision criterion. These criterion shifts are such as to account for the

lower false alarm rate for low-frequency words used as distractors or lures, as compared with the false alarm rate for high-frequency lures. The other class of mirror effect occurs with certain stimulus manipulations (e.g., word frequency), manipulations that presumably affect memory processes. Stretch and Wixted explain mirror effects in this latter instance as being due to a fixed criterion, with the difference in false alarm rates for the two lure distributions being due to their differing on the strength-of-evidence axis. High-frequency lures are judged as more familiar than low-frequency word lures because they are more likely to be activated by associative spread during list presentation.

Now, if our stimulus distinctiveness/memorability manipulation may be seen as analogous to the word frequency manipulation, our mirror effects should have been due to stimulus/memory-based processes, rather than to decision- or criterion-based ones. This sort of effect is what Stretch and Wixted (1998) refer to as a Type II mirror effect. That is, the decidedly lower false alarm rates for caricatured lures than for veridical lures in Experiments 1 and 2 presumably were due to the lower familiarity of the caricatured lures: The mean of their theoretical distribution fell to the left of that for the veridical lures on the underlying strength-of-evidence axis.[7] On the other hand, the distribution of target faces studied in their caricatured form should have been displaced to the right of that for target faces studied in their veridical form. The encoding distinctiveness added by the caricature manipulation increased the probability that the memory representation established during study matched the same target face encountered at test.

However, the mirror effects generated by our observers were not pure cases of a Type II mirror effect. Superimposed upon the aforementioned relative displacements of hit and false alarm distributions was the presence of three different response criteria in Experiment 1 and two different criteria in Experiment 2. As was previously discussed, these criterion differences associated with our stimulus manipulations were independent of the memory-based effects. Nevertheless, in both experiments, the observers appeared to engage in within-list criterion shifts (more conservative) whenever target faces were encountered in caricatured format but had been studied in their veridical form. Although within-list criterion shifting has been rarely noted in the word recognition literature, it is not without precedent (Wixted, 1992).

We would be remiss if we did not raise a possible caveat to the generalizability of these results. It could be objected that our tasks confound face recognition memory and image/picture recognition. Consider first the task in Experiments 1 and 2. Clearly, the target faces in the two crossover conditions of these experiments represent a test of face recognition *qua face*. Here, each of the 16 target faces at test is the same face, but not the same image, as that studied. The 8 faces in each of the other two cells of

our experimental design, strictly speaking, do confound face recognition memory and image recognition when they are presented at test, in that each is the same face and the same image as that studied. However, the logic of our experimental design required that these faces not have their studied images altered by the distinctiveness manipulation. Perhaps we could have had all our target faces presented at test in a different view than that encountered at study. Given the difficulty of recognizing 3-D laser scans of unfamiliar faces in general and the complexity added by our distinctiveness manipulation, however, we did not wish to risk the possibility of floor effects by making the task too difficult. Furthermore, these 16 unaltered target faces only represented 25% of the list of faces presented at test. In the context of 50% *new* faces, presented at two different distinctiveness levels, and 25% of the faces being the same face but in an image of altered distinctiveness, the faces unaltered between study and test might be viewed as simply contributing to the overall complexity of our face recognition task. In brief, we would argue that this confound, one contained in many face recognition memory studies, does not represent a serious limitation on the generalizability of our results.

In summary, we believe that our results lead to a better understanding of how facial caricature affects hits and false alarms, the major performance components of recognition memory. First, caricature increases the hit rate by increasing the effectiveness of encoding processing. The memorability component of face typicality is clearly implicated in effectuating this particular improvement. There is little evidence, on the other hand, for caricature's acting to increase the hit rate by providing more distinctive retrieval cues at test. Second, caricature also acts independently to lower the false alarm rate for caricatured *new* faces. Here, the lower false alarm rate for caricatured lures than for veridical lures presumably is due to our caricature manipulation's also decreasing the general or context-free familiarity of faces. Third, we have shown that caricature is at least as effective in increasing the hit rate as it is in decreasing the false alarm rate.

Finally, particularly at moderate levels of memorability enhancement, we have demonstrated what might be called a *reverse caricature effect*. Study of faces in caricature leads to a better recognition of veridical faces than does study of these same veridical faces. In the present experiments, this was manifested by significantly greater A' scores for 13.5/10.0 and 17.0/10.0 faces than for 10.0/10.0 faces. As was noted earlier (note 6), this phenomenon has potential application to the eyewitness situation, but it also would seem to go directly to the issue of how faces are represented in memory.

REFERENCES

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, **12**, 219-228.

Benson, P. J., & Perrett, D. I. (1991). Perception and the recognition

of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, **3**, 105-135.

BRENNAN, S. E. (1985). Caricature generator: Dynamic exaggeration of faces by computer. *Leonardo*, **18**, 170-178.

BROWN, J., LEWIS, V. J., & MONK, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology*, **29**, 461-473.

BRUCE, V., HEALEY, P., BURTON, M., DOYLE, T., COOMBES, A., & LINNEY, A. (1991). Recognising facial surfaces. *Perception*, **20**, 755-769.

COHEN, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

COHEN, J., MacWHINNEY, B., FLATT, M., & PROVOST, J. (1993). PsyScope: An integrative graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, **25**, 257-271.

DEFFENBACHER, K. A., VETTER, T., JOHANSON, J., & O'TOOLE, A. J. (1998). Facial aging, attractiveness, and distinctiveness. *Perception*, **27**, 1233-1243.

DUDA, R. O., & HART, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

ELLIS, H. D., SHEPHERD, J. W., GIBLING, F., & SHEPHERD, J. (1988). Stimulus factors in face learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues. Vol. 1: Memory in everyday life* (pp. 136-143). Chichester, U.K.: Wiley.

FIORENTINI, A., MAFFEI, L., & SANDINI, G. (1983). The role of high spatial frequencies in face perception. *Perception*, **12**, 195-201.

GLANZER, M., & ADAMS, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, **13**, 8-20.

GRIER, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **75**, 424-429.

HOSIE, J. A., & MILNE, A. B. (1996). The effect of experimental design on memory for typical and distinctive faces. *Memory*, **4**, 175-197.

LIGHT, L. L., KAYRA-STUART, F., & HOLLANDER, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 212-228.

MANDLER, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, **87**, 252-271.

O'TOOLE, A. J., DEFFENBACHER, K. A., VALENTIN, D., & ABDI, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, **22**, 208-224.

O'TOOLE, A. J., MILLWARD, R. B., & ANDERSON, J. A. (1988). A physical system approach to recognition memory for spatially transformed faces. *Neural Networks*, **1**, 179-199.

O'TOOLE, A. J., VETTER, T., VOLZ, H., & SALTER, E. M. (1997). Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age. *Perception*, **26**, 719-732.

RHODES, G., BRENNAN, S., & CAREY, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, **19**, 473-497.

RHODES, G., & TREMEWAN, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science*, **7**, 105-110.

SHAPIRO, P. N., & PENROD, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, **100**, 139-156.

STEVENAGE, S. V. (1995). Can caricatures really produce distinctiveness effects? *British Journal of Psychology*, **86**, 127-146.

STRETCH, V., & WIXTED, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1379-1396.

TULVING, E., & THOMSON, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 352-373.

VALENTINE, T., & BRUCE, V. (1986). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology*, **40**, 300-305.

VOKEY, J. R., & READ, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, **20**, 291-302.

VOKEY, J. R., & READ, J. D. (1995). Memorability, familiarity and categorical structure in the recognition of faces. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition* (pp. 113-137). London: Routledge.

WIXTED, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 681-690.

NOTES

1. It is clear that one cannot manipulate underlying cognitive processes directly. One can only bias the use of different processes. Nevertheless, in proposing to manipulate directly the degree of distinctiveness/memorability of faces at study and test, we believe we have linked our stimulus manipulation about as closely as is possible to the encoding processing occurring during study and the retrieval processing occurring at test. It is also clear that neither the study task nor the test task taps a single underlying cognitive process. For instance, estimates of the effectiveness of stimulus manipulations that bias encoding processing during study are affected by the nature of manipulations that bias retrieval processing during test: The effect of enhanced encoding can only be shown by how it changes the presumed efficiency of retrieval processing.

2. Deffenbacher et al. (1998) also showed that the distance of a face from the average face was related to its attractiveness and apparent age. Increases in distance from the average face resulted in increased memorability, increased ratings of distinctiveness, increased estimates of apparent age, and decreased ratings of attractiveness.

3. Given that the present experiments make use of 3-D surface models only, concerns might be raised concerning their ecological validity and the generalizability of results obtained with them. Certainly, observers recognize 3-D laser scans of unfamiliar faces much less well than photographic images of them (see, e.g., Bruce et al., 1991). One might therefore worry that the underlying recognition processes used by our observers might be rather different from those used to recognize photographic images, for instance. This concern can be at least partly allayed, in that Deffenbacher et al.'s (1998) stimuli were the same 3-D laser scanned faces as those viewed by our present observers and Deffenbacher et al. replicated several common results that have been obtained in the literature on recognition memory with two-dimensional face images (e.g., Rhodes & Tremewan, 1996; Stevenage, 1995).

4. It should be noted that the distinctiveness manipulation employed here was in the context of a within-subjects, mixed-list design (cf. Hosie & Milne, 1996).

5. This most straightforward measure of effect size is computed as $F/(F + \text{degrees of freedom associated with the error term})$.

6. One implication of this result is that it may be possible to increase the accuracy of face recognition for relatively unknown faces, when law enforcement officials search for missing children or suspects, for example. If caricature enhancements can be applied to a photograph, later recognition of the normal, unenhanced face should be improved.

7. The phenomenon of the caricatured lures' having less general familiarity than the uncaricatured lures is evidence that our manipulation of the face memorability component of face typicality may not have been exclusive of corresponding variations in general or context-free familiarity. By lengthening the vector of a given face and thereby moving it farther from the center of the face space, we have not only increased its distinctiveness and memorability (Deffenbacher et al., 1998) but have also apparently decreased its general familiarity. Thus, part of our mirror effect is due to the general familiarity component of face typicality.