Massey University Library. Thesis Copyright Form

Title of thesis: The Effect of Retention Interval and Target-Decoy Similarity on Facial Recognition

(1) (a) I give permission for my thesis to be made available to readers in the Massey University Library under conditions determined by the Librarian.

~~(b) I do not wish my thesis to be made available to readers without my written consent for ______ months.~~

(2) (a) I agree that my thesis, or a copy, may be sent to another institution under conditions determined by the Librarian.

~~(b) I do not wish my thesis, or a copy, to be sent to another institution without my written consent for ______ months.~~

(3) (a) I agree that my thesis may be copied for Library use.

~~(b) I do not wish my thesis to be copied for Library use for ______ months.~~

Signed W. Rockel / Rockel

Date 1/3/91

The copyright of this thesis belongs to the author. Readers must sign their name in the space below to show that they recognise this. They are asked to add their permanent address.

NAME AND ADDRESS

DATE

THE EFFECT OF RETENTION INTERVAL AND TARGET - DECOY SIMILARITY ON FACIAL RECOGNITION

A thesis presented in partial fulfilment
of the requirements for the degree
of Master of Arts in Psychology
at Massey University.

Warren Rockel

1991

ACKNOWLEDGEMENT

I would very much like to thank my thesis supervisor, John Podd, for his expert guidance in helping me through all the stages involved in the preparation of this thesis.

ABSTRACT

This research was an attempt to resolve the inconsistent results for the effect of delay interval on facial recognition. The theory tested was that the degree of target / decoy similarity may act either to enhance or to diminish the effect of delay primarily by influencing false alarm rates. The first experiment used a novel method to scale the 80 faces along the dimension of similarity. The results showed that the method used was reasonably successful in ordering the faces along the similarity dimension. It enabled the use of four sets of 20 faces as either low or high similarity decoy and target sets in a second experiment aimed at testing the proposed theory. It was predicted that high target / decoy similarity would result in a greater effect of delay than low target / decoy similarity. Six groups of 15 subjects completed a standard face recognition experiment which crossed 0, 1 and 21 days delay with high and low similarity target / decoy sets. The results showed a main effect for similarity, but, surprisingly, no main effect for delay. Nor was there the predicted interaction between similarity and delay for false alarms. The failure of the second experiment to test adequately the theory, and reasons for failure are discussed, along with the importance of the link between similarity and delay.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Overview

As humans, the faces of others of our species are for most of us the primary means of distinguishing one person from another. Ellis (1981) wrote that "No other object in the visual world is quite so important to us as the human face."(p.1). It is by their facial characteristics that we learn to recognize friends, relatives, and acquaintances. As social animals, this ability is extremely important if we are to interact and communicate effectively with others of our kind. So our ability to remember and recognize faces is of prime importance to our adequate functioning in everyday life.

The field of facial recognition research is one aspect of the cognitive approach to face processing, which studies how it is that we perceive and remember faces. In the real world outside the laboratory we recognize familiar faces daily. For instance, as we pass someone in the street, by a glance at their facial features we are able to identify that person as someone that we know. We make a judgement as to whether or not the face that we are looking at is that of so-and-so.

For controlled research, it is necessary for experimenters to define what exactly they are investigating. Thomson (1986) distinguishes four different definitions of face recognition:

a) The observer knows that a particular shape or form is that of a face;

b) The observer knows that a particular face has been seen before;

c) The same as b), but the observer knows that the face was seen before at a particular time or place; and

d) The observer knows the name or identity of the face.

For the purposes of the present study, face recognition is considered to be whether the observer knows that a particular face has been seen before. The observer is not required to identify a face by name, nor to state where or when the face was seen before. Recall of faces is a different process, which requires that the details of a particular face be retrieved from memory when the face to be recalled is not actually present. This phenomenon is not a common occurrence in our everyday lives, so is accordingly little studied.

An obvious practical use of face recognition research is for investigating the accuracy of eyewitness evidence. In a criminal trial it is crucial that if an identification on the basis of a suspect's face is to be made, that it can be done accurately. The injustices that can occur after the misidentification of a suspect have been well documented in the literature (e.g., see Yarmey, 1979b; and Shepherd, Ellis, & Davies, 1982). In a police lineup or a criminal trial there is great motivation on the part of a victim of a crime or accident to remember the face of their assailant. And here, too, the complexity and individuality of a face is critical.

It is important to note that eyewitnesses to a crime or accident are most often asked to try to recognize an offender when some time has elapsed since the incident occurred. Just how long after the initial viewing of a face is an observer able to state accurately whether or not that face has been seen before? So it is vital that we know not only the circumstances under which eyewitness testimony may be deemed reliable, but more specifically, for exactly what period of time after the event can a face be accurately retrieved from memory. Hence, the retention interval between study and test phases is one of the most important variables in facial recognition research.

Although common sense might imply that one would expect subjects' memory for previously seen faces to deteriorate over time, research to date portrays a confused picture. Deffenbacher (1986) states that of the 33 studies of forgetting that he examined, roughly half yielded no statistically reliable effect of retention interval. Goldstein and Chance (1981) remark that although the psychology literature is replete with studies demonstrating that forgetting occurs over time, there has been a lack of systematic laboratory research on the effect of delay on facial recognition. Even in the nine years since Goldstein et al. made that comment, very little in the way of systematic research has been carried out. So it seems a fair supposition that there may be other variables interacting with delay to produce these confusing results, and a systematic investigation into what exactly these variables might be is long overdue.

Podd (1990) has suggested that an obvious variable is the degree of similarity between targets and distractors. Davies, Shepherd, and Ellis (1979) have pointed out that in the bulk of face research, target and distractor faces have been selected at random, and systematic study of relative similarity between targets and decoys has been neglected. Davies et al. found that the degree of similarity had an effect on recognition performance in their study.

Shepherd et al. (1982) report the Revised Scottish Guidelines for the composition of identification parades (these may reasonably be taken as representative of practices in Police forces elsewhere) which stress the importance of placing the accused "beside persons of similar age, height, dress and general appearance." (p.133). These guidelines merely codify what has long been known regarding lineup composition: That putting a suspect or target individual in line with decoys who are physically dissimilar to that suspect results in a biased lineup. The suspect "sticks out like a sore thumb", because the witness is given no real choice.

So, there are some indications that the mixed results of previous delay studies could be due to a lack of control over the degree of similarity between targets and distractors. The major aim of this study was to investigate the relationship between delay and similarity.

Typical Recognition Study

In some studies, trying to present as realistic a situation as possible, live crime scenarios are played out in front of an unsuspecting audience. Subsequently they are asked to play the role of witnesses to the crime, in identifying suspects, rating their degree of confidence in their choice, and so on, as happens in actual police investigations (e.g., Buckhout, Alper, Chern, Silverberg, & Slomovits, 1974; Egan, Pittner, & Goldstein, 1977). But not all researchers go to such lengths in attempting to emulate every action that the witness to a crime or incident goes through. More often, the process is reduced to the fundamental act of subjects being shown a face or faces in a laboratory setting, then undergoing a standard recognition test, as described below.

A typical face recognition study involves presenting subjects with a number of photographs of faces, usually in the form of slides presented sequentially. The photographs usually have been black and white, and show only a full-frontal view of the face. This is the "study phase" (or "inspection phase"), and generally subjects are informed that they should pay close attention to the faces (known as targets), as they shall later be requested to attempt to recognize them. However, Courtois and Mueller (1981) found that it made little difference to the results whether or not subjects were told that a recognition test would follow.

In the "recognition phase" (or "test phase"), which can take place either immediately or after a delay, the subjects are shown the target faces again, this time randomly interspersed with other faces, called distractors or decoys, which they have not seen before. Subjects are asked to indicate which of the faces they think they have seen before, by rating them as either old (previously seen) or new (not seen before), often giving the level of confidence in their decision also. The ratio of targets to distractors used in different studies varies a great deal. Laughery, Fessler, Lenorovitz, and Yoblick (1974) used just one target to 149 distractors in their recognition test. However, more commonly a ratio of between 1 : 2 and 1 : 4 is used. Shapiro and Penrod (1986), in their meta-analysis of facial identification studies, found a mean of 22 targets shown at study and recognition phases, with a mean of 40 decoys in the recognition test.

The faces used are either of males only, or of males and females; seldom are only female faces used. Most studies use only white (Caucasian) faces. The length of delay between study and recognition phases varies greatly, with many studies using several different retention intervals for comparison, as well as an immediate test as a control. Deffenbacher (1986) reported a "vast range" of retention intervals tested in the literature on laboratory studies of face recognition, from "one minute to 350 days" (p.63). Shapiro and Penrod (1986) found a mean delay of 4.5 days, with a standard deviation of 21 days.

The present study is limited in that it is a laboratory study of face recognition. The faces used as stimuli are still photographs, and show only a full frontal view. It is hoped that in spite of this simplification in the present study and others in the literature, the subject's task remains an adequate representation of what takes place in the real world, and is generalizable to it.

Signal Detection Theory Measures

Most relatively recent studies make use of Signal Detection Theory (SDT) measures to determine performance in face recognition. In the present study, four of the most common SDT measures were used: hits, false alarms, d', and A_g. The following discussion of these measures is largely drawn from Banks (1970) and McNicol (1972).

In applying SDT to facial recognition, the memory trace is considered as a signal which the subject must detect. SDT is used to separate the truly retention-based aspects of memory performance from the decision aspects (for instance, subjects may appear to be insensitive because they are extremely cautious and only report signals they are certain of). The subject is required to make one of two possible responses to each stimulus, according to whether he or she can detect a memory trace for it: "yes - this is an old item," or "no - this is a new item." Thus, hits occur when the subject gives an "old" response, given that the stimulus was seen before.

A false alarm is when the subject states that a stimulus is old, when in fact it is new.

Hits and false alarms, collected under varying degrees of decision bias, can be plotted against each other to yield a receiver-operating-characteristic (ROC) curve. A_g is the area underneath this curve, being a measure of observer sensitivity, independent of the decision criterion. Like A_g, d′ is also a criterion-free index of recognizability. It is defined as the z-score of the false alarm rate minus the z-score of the hit rate. But unlike A_g, d′ assumes underlying normal-normal equal variance distributions.

According to SDT, all points on an ROC curve represent equivalent retention. They differ only in the degree of caution shown by the subject. A cautious subject may score fewer hits, but also gets fewer false alarms, and, likewise, a lax subject produces more hits but also more false alarms. Some researchers have reported the results of their studies in terms of hits or false alarms alone. This practice of reporting one or other in isolation may be misleading, because either can vary as a result of changes in response bias, and may not in fact indicate that there has been a change in recognition accuracy.

Review of Delay Literature

The aim of this review of delay studies is to demonstrate that not all studies using delay show that recognition accuracy for faces deteriorates with time. I shall give a brief summary of each of the main studies in the literature, discussing first those studies which do show an effect for delay, followed by those which do not.

The retention intervals used in facial recognition studies (excluding immediate tests used to obtain baseline measures) have ranged from a few seconds (e.g., Walker-Smith, 1978) to a study-test interval of 57 years as used by Bahrick, Bahrick, and Wittlinger (1975). Studies such as the latter are extremely rare, due to the obvious difficulties that the great length of time involves. The Bahrick et al. study is the one often cited to show the phenomenal durability of memory for familiar faces over very long time periods.

There are several studies that have used very short delays (only seconds or minutes in length), such as Wallace, Coltheart, and Forster (1970) and Read (1974). However, I shall not discuss these studies because they mainly investigate the phenomenon of reminiscence, in which recognition accuracy has been shown to increase over these extremely short time intervals. This involves memory consolidation, which is a function of the encoding of stimuli in memory; a different process to the retrieval of memories after a more substantial delay.

With retention intervals of between two weeks and 57 years, Bahrick et al. (1975) showed their subjects graduation photographs of high school students, and asked them to pick out the one who had been their classmate from four distractors. They found that accuracy was very high, with around 90% correct recognition even after 34 years. The lowest performance was still a surprisingly high 73% for the group who graduated on average 48 years earlier. However, this study was not a conventional laboratory study, and used faces of people once well known to the subjects, rather than briefly-seen strangers. This makes comparisons with other studies difficult.

A group at the forefront of facial recognition research has been the team of Davies, Ellis, and Shepherd (e.g., see Davies, Ellis, & Shepherd, 1981). A comparatively early study was done by Shepherd and Ellis (1973) on the effect of attractiveness on recognition memory for faces. They hypothesized that subjects' performance on a recognition test would be poorer for faces of a neutral level of attractiveness than for highly attractive or unattractive faces, and that this effect increases with time. Unusually for a study in this field, female faces were used.

Eighteen males and 18 females were shown 27 colour slides and asked to memorize them, as they would be shown some of the slides again and asked to recognize them. As this was a repeated-measures design, with the same subjects used for recognition tests at each of the three delay intervals, a different third of the target set was shown at each test. Two delay periods were used; six days and 35 days, as well as an immediate recognition test as a control. The recognition test consisted of a dual presentation of target with distractor, with subjects asked to indicate which was the previously-shown slide of each pair.

Significant results were obtained not only for the hypothesized interaction between attractiveness and testing interval, but also for the main effect of testing interval. The number of stimuli correctly recognized (out of nine, chance = 4.5) declined from 8.05 (immediate test) to 7.28 (6 days) to 6.42 (35 days). One interpretation that Shepherd and Ellis (1973) offer for the observed interaction, is that faces high or low in attractiveness may be more "distinctive" than neutral faces, making them more memorable.

Egan et al. (1977) investigated the differential accuracy of eyewitness identification using photographs versus live models. Their procedure was designed to resemble the process gone through by the witness to a crime. Subjects were 50 male and 36 female university students who viewed two live male targets dressed alike. The test phase took place after a retention interval of either 2, 21, or 56 days. Half the subjects saw a corporeal lineup, the others seeing photographs of one of the targets with four decoys. Two photographs (one full-length and one full-face) were shown of each person in the lineup. It was found that live targets were correctly identified 98% of the time, compared with only an 85% hit rate for photographs. Interestingly, although there was no effect of delay upon hit rate, false alarms increased from 48% to 93% over the 56 day period (p < .10).

Lipton's (1977) study on the psychology of eyewitness testimony took the next step, in simulating a courtroom setting rather than a police lineup. Without any specific instructions being given, subjects saw a film of a crime. Then either immediately or after one week, subjects were directed into another room for the "trial" in which they were questioned by a mock lawyer on details in the film. After the week long delay, accuracy of details remembered was 4.3% lower than the immediate test, and the quantity of information was 18% lower.

Davies, Ellis, and Shepherd (1978), as part of their investigation on the influence of delay on "Photofit" construction, also measured ability to select a target from a sequence of 30 faces. This time the slides used were of men, of roughly the same age as the target, and similarly dressed. Twenty subjects were in each delay period: three weeks, and within 48 hours (designated as "immediate" test). Results showed that subjects had a significant tendency to score lower on the recognition test after three weeks.

Continuing the work done by Shepherd and Ellis (1973), Yarmey (1979a) investigated the effects of attractiveness, feature saliency and liking on memory for faces. Thirty faces of males and females were studied by 126 male and female subjects, who were required to judge for each face either its degree of attractiveness, or distinctiveness, or likeability, as well as to memorize the faces for a later recognition test. One third of subjects went into each delay condition: immediate test, seven day retention interval, or 30 day interval. Results were much as Shepherd et al. found: Recognition performance decreased over time, and faces rated as high or low on a given dimension were more easily recognized than medium or "neutral" faces.

Krouse (1981) examined the effects of pose, pose change, and delay on face recognition performance. Commenting on the mixed results of previous studies, she suggests that length of delay alone does not entirely account for the presence or absence of significant differences. Using 16 targets and 48 distractors, Krouse found a decrease in overall recognition accuracy after a 2 - 3 day delay. On the basis of her results, Krouse suggests that faces are better recognized in some poses than in others, with the three-quarter pose yielding more accurate recognition performance than the traditional full-face view used in police "mug-shots" (and also in most face recognition research).

Courtois and Mueller (1981) investigated target and distractor typicality in facial recognition as an examination of bias in lineup composition. In addition to varying delay, they studied the distinctiveness of target and decoy faces. (This aspect of their study is discussed in further detail below, under similarity.) Ninety-six male and 96 female subjects viewed 10 faces (five male and five female), facing a recognition test either immediately, after two days, or after 28 days. They saw 10 arrays of four faces, each containing a target slide which they had to choose. Courtois et al. found significant decreases in recognition between one minute and two days and also between two days and 28 days.

Deffenbacher, Carr, and Leu (1981) are among an increasing number of researchers who make use of SDT measures. Like Barkowitz and Brigham (1982, discussed below) they had a distractor-to-target ratio of 2 : 1 for the immediate test, but they also had a two week delay in which the ratio was 3 : 1. (Although, in theory, not affecting recognizability, this does affect response bias. The more that distractors outnumber targets, the more likely a subject is to say that a given face is a distractor.) Deffenbacher et al.'s examination of memory for words, pictures and faces varied not only delay but also gauged the effects of interference and reminiscence.

Half the faces were of males, half of females. Facial hair, glasses, jewellery and unusual clothing were eliminated, and backgrounds were neutral. Cell sizes ranged between 10 and 17. Subjects were told to study the 21 stimuli carefully, as a recognition test would follow. Before the recognition test, subjects were informed of the probability of a target occurring in each item. The resulting average hit rate for faces was .81 for the immediate test group, and dropped to .68 for the two week delay group. Giving hit rate alone is dangerous because there is no way of being certain whether the results are due to a change in bias or a change in recognizability, or if both are contributing. However, the authors also gave results for false alarms old (FAO, distractors used in an earlier part of the experiment) and false alarms new (FAN, distractors not seen before). The former increased from .10 to .14 over the two weeks, while FAN increased from .06 to .10. Thus it is possible to say that delay did affect recognizability, and these results cannot be said to be solely due to response bias.

Unlike the Davies et al. (1978) study, Barkowitz and Brigham (1982) had clearly delineated immediate and 48 hour delay periods, as well as a seven day delay, in their study on own-race bias, incentive and time delay. A mixture of black and white, male and female subjects (10 subjects per cell) viewed 24 target slides (six from each of the four sex-race categories). Faces with unusual expressions, glasses and excessive jewellery were eliminated from the set. As is the norm in studies where a recognition test is to take place, subjects were told that they were going to be shown a series of slides which they would later be asked to identify. Subjects were asked to mark on a response sheet whether each of the 72 slides they saw in turn was "old" or "new".

Although no effect of delay was evident in the results for female faces, recognition of male faces did deteriorate over time. An interesting additional finding for Barkowitz and Brigham (1982) was that subjects adopted more lax decision criteria as time went on. This led them to suggest that their subjects were "filling in the gaps" in their memory as details faded over time.

One of the more recent studies on retention interval and face recognition was carried out by Chance and Goldstein (1987) on response latency measures, comparing Caucasian faces with the more difficult faces of Japanese models. Subjects were 59 male and female Caucasians who each saw 16 Caucasian and 16 Japanese faces in the study phase. The recognition test took place either immediately, or after two or seven days. Eight of the original slides from each target set was shown with 24 distractors of each race. No effect for delay on hit rate was found, but interestingly, false alarms increased for both Caucasian and Japanese faces with the longer intervals, presumably showing real forgetting.

Podd (1990) examined the effects of memory load and delay on recognition of photofit faces. He intended his study to be part of a series examining the effects of retention interval on a range of variables, pointing out that the reasons for the mixed results evident in the literature on the effects of delay remain inadequately researched and are as yet unexplained. The 29 male and 61 female subjects saw either 20, 35, or 50 faces at study with the same number of distractors at recognition.

In the recognition phase, which took place either 10 minutes, one week or two weeks later, subjects were informed that equal numbers of old and new faces would be shown in random order, with an equal probability on each trial of an old or a new face appearing. A four-point rating scale was used: (1) very sure old, (2) fairly sure old, (3) fairly sure new, (4) very sure new. The results revealed that increasing memory load decreased accuracy largely by lowering the hit rate. Conversely, delay affected recognition accuracy mainly by increasing false alarms (as found by Chance and Goldstein, 1987).

In summary, the above 12 studies show clearly the expected drop in recognition accuracy with the passage of time. This effect has been found using delays as short as two days (a commonly used period) and as long as 57 years if the Bahrick et al. (1975) study is counted. Faces have been in the form of photographs, slides, photofit compositions, and live lineups. There are numerous other differences between these studies, yet they all manage to yield an effect for delay. In spite of there being this quantity of strong evidence for an effect of delay on facial recognition rates, the literature also contains a number of instances where delay seems to have had no effect.

Active in the facial research field for some time have been the Goldstein and Chance group. Goldstein and Chance (1970) studied visual recognition memory for complex configurations to test their hypothesis that human faces are a unique and overlearned form of stimuli, which they compared with ink-blots and photographs of snow crystals. The faces used were full-face photographs of women, standardised by holding constant head position and camera angle. Cues such as jewellery, glasses, and any facial blemishes were eliminated from the set of stimuli.

Subjects in the faces condition were told to pay attention to the stimuli, and that their recognition would be tested. Half of these subjects were tested for recognition immediately, the others 48 hours later. (There were 13 males and 13 females in each group.) They were asked to pick out from 84 faces the 14 they had seen before. Apart from showing that faces were better recognized than ink-blots and snow crystals, the results with regard to faces were confusing. Contrary to what one would expect, responses to faces did not show an effect of delay, even though there was a 17% decrement for each of the other two forms of stimuli.

A similar study was done by Chance, Goldstein, and McBride (1975). A part of their series of experiments on differential experience and recognition memory for faces, investigating race effects on recognition, was a partial replication of the previously discussed study (Goldstein and Chance, 1970). A difference was that the stimulus faces were half male and half female. Results were as above - no effect of delay after 48 hours.

An example of a study using a very short delay is Yarmey (1971) who looked at recognition memory for familiar "public" faces, and the effects of orientation and delay. The 40 male and 40 female subjects inspected 80 "flash cards" consisting of black-and-white photographs of 20 familiar famous male people, 20 unfamiliar males, 20 faces of dogs and 20 photographs of buildings. In the recognition test, half of these study photographs were replaced with new pictures. Half of the subjects saw stimuli rotated 180 degrees. Forty of the subjects had an immediate recognition test, the other 40 had a 20 minute delay. The rotation resulted in greater recognition errors for both familiar and unfamiliar human faces than for the other stimuli. No effect of delay was evident.

Laughery et al. (1974) did a series of studies on time delay and similarity effects on facial recognition. Their study on delay is unusual in that they used so many distractors (149) to just one target. The target was shown either in slide form in four different photographed positions, or in a one-minute film clip playing a significant role. The recognition test showed all decoys and the target in front bust views. Retention intervals of four minutes, 30 minutes, one hour, four hours, one day and one week were used. Number of subjects per condition ranged from nine to 14.

No effect of delay on recognition performance was found. However, this may well have been due to subjects' response bias. Presented with 150 stimuli, of which just one is the target, chances are greatly on the side of subjects choosing a distractor in error, rather than a target if they are unsure of which it is. Indeed, Laughery et al. (1974) conclude that their series of studies (the others are discussed below in the review of the similarity literature) "clearly indicate that the important variable in facial recognition is the number and similarity of the faces preceding the target picture in the search series and not the amount of time elapsed" (p.496). An evocative suggestion!

To summarise, the above four studies do not show an effect for delay. Although apparently using procedures similar to some of the previous 12 studies, for an as yet unexplained reason these researchers could not duplicate the findings of the other studies. Some have suggested that it was not delay which caused the drop in recognition in the other 12 studies (e.g., see above, Laughery et al. 1974). However, the weight of evidence strongly points to delay playing an important role in recognition accuracy.

As this review has shown, there is a problem of conflicting results which has yet to be satisfactorily explained. As proposed earlier, the hypothesis which the present study aims to investigate is that a lack of adequate control over the degree of similarity between target and distractor faces may account for the confused findings demonstrated. This theory will be further outlined below.

Theories

There has been very little theoretical development in face research to account for the above conflicting findings, with most researchers concentrating on the more practical implications. This has left many questions unanswered. Two of the most enduring enigmas will be considered here: a) How are faces stored in memory?; and b) Would one expect recognition rate to be affected by lengthening delay? It is important that face recognition research be designed to solve such theoretical problems.

Concerning the first question, there are two general classes of theory regarding storage of faces: the holistic approach, and the piecemeal approach (see Sergent, 1984). Put briefly, the principal difference between these two schools of thought is that holders of the holistic viewpoint propose that faces are stored in memory as a whole complete unit. On the other hand, followers of the piecemeal approach believe faces to be stored as a collection of separate features.

As is the case with the delay literature, evidence is inconclusive as to which theory is most likely. Ellis (1986) mentions this as one of "ten questions in need of answers" in face processing research. In the past it has been a widespread assumption that faces are processed as a complete whole or Gestalt. How can such a uniquely individual stimulus as a human face, which is normally seen as a whole, be considered as an ensemble of parts? But, as Sergent (1984) points out, there is actually very little evidence in the literature which supports the holistic view.

In fact, lately, evidence has come to hand which has supported the alternative view, most notably the observation that certain features of the face are recalled more accurately than others. Studies of feature saliency have shown that generally it is the upper facial features (eyes, forehead, hair) that are better remembered. In Shepherd, Davies, and Ellis' (1981) review, four out of six studies found this advantage, and across all the studies they examined, hair emerged as the most important cue.

Sergent's (1984) investigation into component (piecemeal) and configural (holistic) processes underlying face perception attempted to draw together both theories. She concluded that "Multidimensional stimuli such as faces have component properties as well as configural properties that emerge from the relationship among the features, and the processing of these stimuli may be based on either or both of these properties." (p.237). Thus, it appears that both piecemeal and holistic processes can operate simultaneously. This theory acknowledges the importance of the relationship between facial features, as well as the features themselves, in giving the face its individual appearance.

Sergent (1984) concedes that some questions remain unanswered by her theory. At this stage it is still difficult to say which of the two processes are most important to us in our everyday lives, and in what particular situations holistic processing takes dominance over piecemeal, and vice versa. It remains impossible to conclude on the basis of the existing evidence whether we store faces in memory as a whole, or as a collection of parts. More theoretically based research needs to be done to answer such problems.

However, a theory can be derived from some of the existing literature which has at least an indirect bearing on the piecemeal vs. holistic controversy. There is a small number of studies (e.g., Egan et al., 1977; Chance & Goldstein, 1987; and Podd, 1990) that show the interesting occurrence of hit rate remaining constant with delay, while false alarms increase. This phenomenon may be significant in providing a partial answer to both the questions set out at the beginning of this section.

To explain the above occurrence, Podd (1990) argues that when recognition rates for high and low similarity target / decoy sets are compared there is no difference in hit rate as faces previously seen are matched to their stored images in memory. False alarm rate suffers with high similarity target / decoy sets because more facial features in the decoy faces match those in target faces. Thus, false alarm rate increases.

According to Podd (1990), delays of up to two weeks have a bigger impact on false alarm rate than hit rate because memory for the target faces remains relatively good over a two week period, so hit rate is unimpaired. On the other hand, false alarm rate increases as parts of the face image stored in memory fade with time. Facial decomposition is not serious enough to cause a drop in hit rate, but false alarm rate is affected because fewer features remain accessible in memory. Therefore, there is a greater likelihood of a decoy face being matched to a stored face. Hence, there is an increase in false alarm rate.

According to this theory, high similarity target / decoy sets will enhance the effect of delay. So, higher false alarm rates will occur with delay, and will be even higher when combined with high target / decoy similarity. This reasoning implies that faces are stored in memory in a piecemeal fashion. As retention interval increases, the features of the target faces retained in memory decrease in number. With fewer facial features in memory to discriminate the target, the number of distractor faces sharing similar remembered features increases, thus raising the possibility of the subject selecting the wrong face, and increasing the false alarm rate. Podd goes on to offer the explanation that if different facial characteristics decay at different rates, and, following a delay, fewer details are available in memory to compare the test faces with, then this effectively increases the perceived degree of similarity between targets and distractors.

Therefore, the question is, do delay and similarity have similar effects? Podd says that they do. If this explanation is correct then predictions can be made for the case where delay and similarity are directly manipulated. Podd predicts a higher false alarm rate for high similarity target / decoy sets, having fewer characteristics available to distinguish between targets and distractors, than for low similarity target / decoy sets. Under the theory it is also expected that delay will further increase false alarm rates, as elaborated above.In addition, if the theory for delay is supported, this will provide indirect support for the piecemeal theory of face processing.

Thus, a major implication of the theory is that high similarity target / decoy sets will yield higher false alarm rates following a delay than low similarity target / decoy sets. In other words, false alarm rate should be highest for high similarity target / decoy sets after a delay and lowest for low similarity target / decoy sets after a zero delay. It is possible, of course, that low similarity sets, or even supposedly high similarity sets, may be so distinctively different with respect to targets and decoys that a delay in the region of two weeks may have little effect. This is a matter that will be further discussed in considering the results of the second Experiment.

Review of Similarity Studies

While there are a relatively large number of studies which vary the length of retention interval in facial recognition, there are almost none which have varied the degree of similarity between target and decoy faces. A probable reason for this dearth of similarity studies is the problem of how to rate faces for similarity before they can be used in an experiment.

Among the first researchers to make an effort to control similarity, and address the problem of trying to define it, were Laughery et al. (1974). They performed a series of studies which varied the presence of similarity between the target and the distractors which appeared in the series before the target, in the context of investigating the effect of target position on recognition performance. As noted above in the discussion of the delay literature, Laughery et al. used 149 distractors to one target. They carried out three studies, each using a different definition of similarity:

a) For each of the decoys a mean score across subjects was obtained that showed the degree to which each decoy was mistaken for a target. Thus, the more often subjects called a particular decoy a target, the more similar that decoy was considered to be to it.

b) The people taking the pictures for the film library made estimates of each model on nine physical characteristics (hair colour, hair length, age, build, eye colour, glasses, moustache, beard, and length and shape of sideburns). For 315 of these pictures of white males, the number of characteristics each pair had in common were counted. High- and low-similarity distractors were then selected from a similarity matrix of these 315 faces. High similarity faces were those which had eight or nine characteristics in common. Low similarity faces had four or less features in common. Four faces were chosen as targets, on the basis that 52 similar distractors and 52 dissimilar distractors could be found for them in the set.

c) A "large number" (p. 494) of subjects (a minimum of nine per pair) rated 222 faces in pairs for similarity. From the obtained matrix, the relative similarity of any two faces could be defined.

For hit rate, Laughery et al. (1974) found significant results for b) and c) and nearly significant results for a), which indicated that the lower similarity decoys produced a higher hit rate. A prediction of higher similarity leading to higher false alarms was only partially supported, with the effect being significant only in a) and half of c).

Egan et al. (1977), discussed above with the literature on delay, held similarity constant by selecting models to look alike. They were all of similar height, weight, and hair colour. No further detail is given of the process used to determine similarity, but it appears to have been a simple screening process - a quick look at each model to determine if they look like those already chosen - rather than a carefully defined system.

The results obtained by Egan et al. (1977) were as predicted by Podd's (1990) theory: with a high degree of similarity between targets and distractors, hit rate remains relatively constant over delay, whereas false alarms increase. These results also reflect the earlier findings of Goldstein et al. (1977), that certain distractor faces attract more than their statistical share of false alarms. Goldstein et al. suggested similarity as one possible reason for this.

Neither of the above studies has brought the field much nearer to finding a valid method for rating similarity, the former because of its weak effect due to the high level of bias inherent in using a 1:149 target:distractor ratio, the latter because of its poorly-defined method. Later research has come further, though.

Cohen and Nodine's (1978) study clearly implicates the use of highly similar distractors as being associated with an increase in false alarms when compared with low similarity distractors. Identikit line drawings of faces were used as stimuli. Each target face was presented in a recognition test with two distractors which differed from it by two, four, or all seven facial features. The false alarms were in the expected direction - the greater the number of features a distractor had in common with its target, the more often it was mistaken for a target. This method of determining similarity is simple and appropriate for standard line drawings, but what about for more realistic representations of the human face, such as photographs?

When faces must be rated on some continuum, such as level of attractiveness, it is possible to obtain valid results by simply asking subjects to give their subjective impressions. For instance, subjects might be shown a series of 100 slides of male faces, for five seconds each, and be requested to rate each face for its degree of attractiveness on a 10-point scale. This is possible because an attribute such as attractiveness has an existence outside of the set of faces being used. Similarity, on the other hand, can only be judged with respect to a particular set of faces, which must be simultaneously considered relative to each other. One can instantly decide whether a given single face is attractive or unattractive; but one cannot judge similarity without some point of comparison.

So to use a subjective method for rating similarity, it is necessary to provide a point of comparison. The ideal way is to devise a system whereby each face is compared with every other face in the set to be used. Davies et al. (1979) did just this in their study, which examined similarity effects in facial recognition by applying hierarchical clustering analysis (HCA) to a similarity matrix based on direct comparisons of physical resemblance.

The models for the photographs were 100 male Caucasians aged between 17 and 55. Cues such as background, facial expression, clothing, facial hair, and glasses were eliminated. Each face occupied about two-thirds of the 8.5 cm x 11.5 cm prints. The subjects (24 males and 24 females aged 18 - 60) were individually given the set of 100 photographs and requested to group them into piles on the basis of physical similarity. The number of piles which subjects chose to use ranged from three to 33. The subsequent groupings were then entered into a 100 x 100 matrix, showing the frequency with which any one face was sorted with another.

These frequency scores, considered to be similarity measures, were analysed by Johnson's HICLUS Program (see Johnson, 1967). The result was a series of clusterings of stimuli at increasing levels of generality. At the lowest level, each face is a cluster on its own. At the next level the two most similar clusters (faces) are merged to form a new cluster; and so the process continues up to the selected level. Davies et al. (1979) chose a level of clustering which provided an adequate number of clusters of sufficient size for use as decoys in the recognition test. This gave them four clusters of 6, 6, 7, and 16 faces respectively. The targets were taken from the largest cluster, with decoys coming from either the same (high similarity) or different (low similarity) clusters. The prediction was that distractors drawn from the same cluster as the targets should be more readily confused with them than decoys from different clusters.

Davies et al.'s (1979) first two experiments sought to establish the validity of the subgroups identified by the clustering program. The authors used a within-groups design to investigate how error rates were affected by the method of presentation of the recognition array (successive vs. simultaneous), the number of targets present (all ten, five, or none), and the style of instructions (stressing a strict vs. a lax criterion). No effect was found for method of presentation, but there was a large increase in errors when no targets were present, and the "strict" criterion instructions resulted in fewer errors. The overall aim of these two experiments was to show that distractor faces from the same cluster as the targets would more likely be wrongly identified as targets than faces from the other three clusters. In fact, common cluster membership accounted for 72% (Expt. I) and 84% (Expt. II) of the false alarms.

Experiment III was a between-subjects design similar to that used by Laughery et al. (1974). The 34 subjects were randomly allocated to either the same or different cluster conditions. The presentation phase was the same for all subjects. They were shown four target slides one at a time, which they were asked to study carefully so as to be able to recognize them later. The faces appeared as roughly double life size, with subjects seated 3 - 4 m from the screen. The recognition phase took place five days later. Half of the subjects saw the four target faces interspersed with the remaining 12 high similarity decoys, while the other half saw them with 12 faces drawn randomly from the other three clusters. Subjects were asked to rate on a five-point scale their certainty as to whether a given face was a target or not.

As expected, similarity acted to increase false alarms, while hit rate remained unimpaired. This finding provides indirect support for the piecemeal, or feature selection, theory of face processing. Misidentifications can be made even when faces are alike on only a small number of characteristics (though certain similarities exist, the faces are far from identical to each other). An interesting additional finding was that the differences between clusters largely consisted of variations along four main dimensions: Hairstyle, face shape, age, and eyes, were the most salient features for the subjects.

The present study also needed to scale the faces for similarity, but the Davies et al. (1979) study did not provide the perfect method for our purposes. We wished to obtain four clusters of equal size (20 faces in each) so that all subjects could see the same numbers of targets and distractors. It was highly unlikely that we would have got this under HCA, so a new system was invented (see Expt I: Method).

Two further studies have investigated an area closely related to similarity; that of distractor "distinctiveness" or "typicality". Light, Kayra-Stuart, and Hollander (1979) examined recognition memory for typical and unusual faces in a series of experiments. In the last of these they had two aims. Firstly, they sought to discover whether the mean inter-item similarity of faces predict typicality judgments. Their second aim was to see whether the lower similarity of unusual faces accounted for their greater recognizability. Subjects rated 30 slides presented in all 435 possible pairs on a 7-point scale, from 1 "not at all similar" to 7 "virtually identical". The photographs were head and neck shots of white male high school seniors all dressed alike and against the same background, but oriented in a variety of directions and wearing a range of facial expressions. Also those with unusual facial characteristics (e.g., glasses) were not excluded. The 14 male subjects also rated facial prototypicality from "least typical"

(1) to "very typical" (7).

The typicality and similarity ratings correlated .81. It therefore appears that there is a strong relationship between typicality and similarity (which is of relevance to the Courtois and Mueller (1981) study, discussed below). Light et al. (1979) also showed that faces similar to other faces are more difficult to recognize - increasing typicality is associated with greater false alarms. This provides support for Podd's theory, as their subjects were more likely to mistakenly identify typical new faces as old than they were for unusual new faces.

Courtois and Mueller (1981) investigated "typicality" or "distinctiveness" of faces, which they consider to be a "more subtle factor" (p. 639) than similarity. In practice, this has much the same effect as similarity because, essentially, both refer to how much targets stand out from the decoys due to physical appearance. The .81 correlation between similarity and typicality ratings obtained by Light et al. (1974) indicates a close relationship between the two. The instructions given to the initial 10 male and 10 female subjects were to rate each picture in terms of how similar it was to their idea of the typical college senior. So, in effect, these subjects must have had in mind an image of the "typical university student" to which they compared each of the faces in terms of their perceived similarity to it.

The hypotheses tested by Courtois and Mueller (1981) were that typical target faces would be difficult to identify when embedded with other typical faces, but their identification would be enhanced when the distractors are atypical. On the other hand, if unusual or distinctive faces have unique features, there would be little difference in recognition rate whether they are shown with typical or atypical decoys.

The faces were of 80 male and 80 female dark-haired Caucasian university students, without beards, moustaches, glasses or earrings. Subjects saw the 10 targets with 30 decoys in the recognition test. The results were as Courtois and Mueller (1981) had predicted. The least accurate recognition occurred when a typical target was paired with typical distractors. The most accurate recognition occurred when a distinctive target was paired with other distinctive faces (i.e., target and distractors were dissimilar). Courtois et al. got a significant overall effect for retention interval. Thus it seems appropriate to take from Courtois et al's results that a lack of control over distinctiveness/similarity could well account for other studies not getting an effect for delay.

In summary, as the review of the delay literature shows, the findings on the effect of delay on facial recognition are confused - some show this effect, others do not. Common sense, and the experiences of our own everyday lives, tells us that although memory has been shown to be remarkably durable, our memories are far from perfect, and accuracy does decrease with time. Until these mixed results are accounted for, it remains difficult to conclude exactly what the effect of delay is.

Part of the problem may be an over-strict adherence to the ".05 rule". Just because an effect is not statistically significant, does not mean the trend is not present. Deffenbacher (1986) found that there was in fact a statistically reliable effect for delay, when averaged across all 33 studies which he examined, even though many of the individual studies failed to yield significant results. It is not yet clear why the non-effect occurs, mainly because there has been a lack of theoretical investigation into likely reasons. One explanation has emerged, which suggests that delay may be confounded by physical similarity between targets and distractors.

Podd (1990) argues that when distractors are used which are similar to the targets, they are more easily confused with the targets than are low similarity distractors, resulting in a higher false alarm rate. Hit rate is comparatively unchanged because the subject's ability to identify targets remains unaffected (see below). Over time, the effect of highly similar target / distractor sets increases because delay itself, in effect, increases target / distractor similarity. One can extrapolate from this to reason that studies obtaining an effect for delay may have used distractors high in similarity to the targets. These are more likely to be confused with each other, so subjects are less easily able to pick out the targets.

Podd's theory provides indirect support for the piecemeal approach to memory for faces, which assumes that faces are stored as a collection of features, rather than as a whole (holistic view). We can reason that target and decoy faces are confused with each other because they have features in common. When a subject looks at a face and is asked to decide whether it is an old or a new face (a target or a distractor), he or she scans their memory "looking" for this face. Because faces are stored as features, the more characteristics that the test face has in common with a previously seen face, the more features the subject finds in memory that are similar to those of the test face. Thus the subject is fooled into thinking he or she has seen this face before, and in mistaking a decoy for a target makes a false alarm. This is more likely to occur if some time has elapsed, and some of the stored features have disappeared, leaving less cues available to discriminate between faces.

Before this theory could be tested, it was necessary to rank the stimulus faces for their similarity. Thus, the first of the two studies reported had the aim of ranking faces for similarity. The second study investigated the relationship between delay and similarity.

EXPERIMENT I

OBTAINING THE SIMILARITY RATINGS

Before investigating the relationship between similarity and delay, the set of photographs to be used first had to be ranked in order of similarity. Because so little research has been done on facial similarity, there was very little guidance in the literature on exactly how to have our subjects sort the set of faces in order to obtain the similarity ratings. The study by Davies et al. (1979) involved the use of a hierarchical clustering program in which each face is clustered with other similar faces, and each cluster in turn is joined with other similar clusters through subsequent increasing levels of generality. No details are provided in the article of the clustering program Davies et al. used, but it seems that it allowed them little control over the relative sizes of the clusters. They had their subjects sort 100 photo's into piles of similar faces, and after HCA obtained four groups ranging in size from six to sixteen faces.

For the purposes of the present study, it was necessary to obtain four equal-sized groups. HCA programs are inadequate for this. The HCA programs within SPSS.X and Genstat (the packages available at Massey) allow the user to halt the program when it has produced enough clusters of sufficient size, but do not have a facility for setting a standard desired size for each cluster.

Other scaling methods proved impractical because of the large number of stimulus faces (80). Torgerson (1958) outlines a Multi-Dimensional Scaling (MDS) technique called the "method of triadic combinations" (p.262), in which all possible triads of the stimuli are presented, and the subject must decide which two of the three are most alike, and which two are most different. However, due to the large number of stimulus faces this would present the subject with a phenomenal task. Torgerson also details the "method of paired comparisons" (p.166), in which each stimulus is compared with every other stimulus, with regard to degree of similarity to a reference point, for instance. This yields the proportion of times that each stimulus is judged greater than any other on that dimension. With 80 stimuli there would be 80(80-1)/2 = 3160 pairs. It is too much to expect a subject to rate that many pairs!

A more practical approach to sorting a comparatively large set of stimuli, as was the case in the present study, emerged from Torgerson's (1958) discussion of Thurstone's "law of categorical judgment." From this law a rank-order procedure is derived, involving many subjects each placing stimuli in rank order with respect to the attribute in question. It was decided that a method along these lines would be best for the requirements of the present study, thereby avoiding placing an unreasonable burden on subjects, with a possible accompanying loss of accuracy due to boredom and fatigue.

To ensure the maximum possible validity of results, a "modified" triadic process was added to the aforementioned procedure. Subjects compared each face with both the anchor face and with each of the previously ranked faces in turn, making a decision where in the ranking that face belonged, inserting it between a face that was more similar to the anchor and the one that was less similar. In our modification, the subjects were forced to narrow their focus to three faces at a time. The end result was a simple rank ordering of the 80 photographs, beginning with the anchor face, with the others following in descending order of similarity to the starting face. After averaging across subjects it was then possible to divide the stimuli into four groups of 20 to obtain the category sets of faces.

Positions 1 to 20 was the target set (including the anchor face as position 1, along with the 19 faces rated as most similar to it). Positions 21 to 40 represented the high similarity distractor set. Positions 41 to 60 represented the unused "separation" set. This set was necessary to ensure that the overlap between high and low similarity decoys was minimal. Positions 61 to 80 represented the low similarity distractors (being those faces rated as least similar to the anchor face). The first two and the latter one of these groups constituted the three categories of faces used in the main study. Equal sized groupings were desired so that all subjects saw the same number of targets and distractors in the main experiment.

METHOD

Subjects

A total of 38 volunteer subjects (18 male and 20 female), aged between 22 and 50, drawn from among the academic and clerical staff and postgraduate students of the Psychology department at Massey University, served as judges for ranking the similarity of the faces.

Stimulus Faces

The colour photographs of 80 male Caucasians were obtained mostly by recruiting undergraduate students, both internal and extramural, from the Massey University campus. A number of people were also solicited from various workplaces around Palmerston North. The models' ages ranged from 18 to 75 years, with the vast majority falling between 18 and 30 years.

Any potential memory cues, apart from the faces themselves, were controlled for by ensuring that none of the models were balding, had any facial hair, earrings, or other adornments or unusual features. Those wearing glasses were asked to remove them for the photograph. All photographs were taken full-face, with the model looking straight into the camera, wearing a neutral expression, against a neutral background, and with identical lighting conditions for each shot. A cape was worn to obscure all clothing.

Materials

The primary materials used in this study were the set of 80 face photographs. These measured 10 x 15 cm. One of the photographs was designated the starting face. This was the same for all subjects, to give a common point of comparison for all the other faces to be related to in terms of similarity. Several tables were joined end-to-end to provide space for spreading out the photographs.

Procedure

Each subject was individually given the set of photographs (which were in a randomised sequence), and instructed to order them on the basis of their physical similarity to the starting face using the modified triadic method, which was explained to them. Subjects were also told that, if they preferred, they could first sort the faces into groupings of similar faces if they felt that made their task any easier.

In addition, it was suggested that if they were unsure as to what to base their similarity judgments on, they could consider the characteristics of face shape, degree of eye roundness, and hair length. (These were found by Davies et al. (1979) to be characteristics frequently used by their subjects in judging the similarity of faces.) Subjects were given a piece of paper listing these three criteria, then left alone to perform the task.

When finished, subjects were debriefed and the order they had put the faces into was noted in terms of the arbitrary catalogue numbers written on the back of each photograph. One subject was dropped from analysis because of suspicion about the validity of his rankings due to the extreme haste with which he performed the task, taking less than half the time of any other subject.

Determining the Final Rank Order

Once a number of subjects (see below) had ordered the faces, their rankings were entered into a Hewlett-Packard Series 300 computer. A purpose-written program then calculated, for all subjects entered up to that point, the number of times that each of the faces was put into each grouping or category, listed category by category (20 faces for each category). Printouts were obtained at intervals (after 3, 10, 20, 30, and 38 subjects had been entered) to see if a pattern was emerging, at which point it should be possible to cease running subjects. As there had been very little change in the pattern of results between 30 and 38 subjects, it was decided that little was to be gained by running further subjects, so the analysis based on 38 subjects was decreed the final one.

The original intention had been to allot each face to a certain rank order position, according to which position the greatest number of subjects had put that particular face into. While this was possible for the majority of faces, a few seemed difficult to categorize with respect to similarity to the starting face. These faces showed wide variability in rank ordering for individual subjects. Therefore, as the most important factor for the aims of the present study was the category which the greatest number of subjects put a face into, rather than its specific rank order within that category, it was decided to calculate the final sequence using the following method.

Taking one face at a time, the number of subjects who had put it into either positions 1-20, 21-40, 41-60, or 61-80 was calculated, these being the positions comprising each of the four categories. The face was put into whichever category it had the greatest number of votes for, irrespective of its specific position within a category.

Not all categories had the same number of faces in them, as there were eight faces which were tied, or very nearly tied, between categories. These faces were allocated to the tied category which was short of its complement of 20.

RESULTS AND DISCUSSION

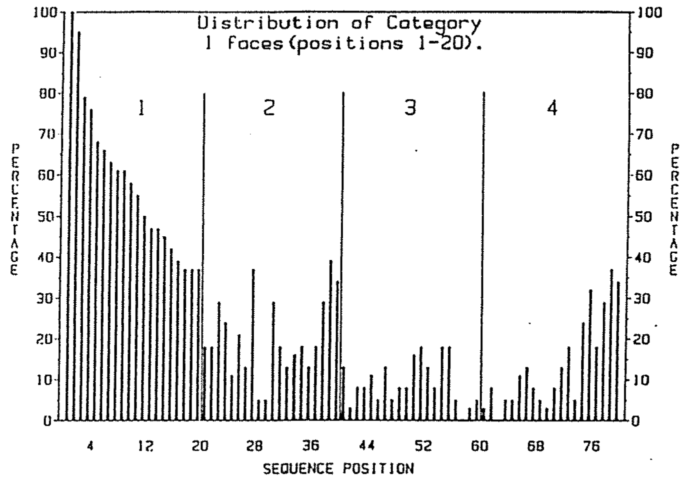
Having divided the 80 faces by the above method into four groups of 20, ranked in terms of similarity, it was important to assess the method's reliability. The question is, did the method produce a noticeable differentiation of similarity across the four categories?

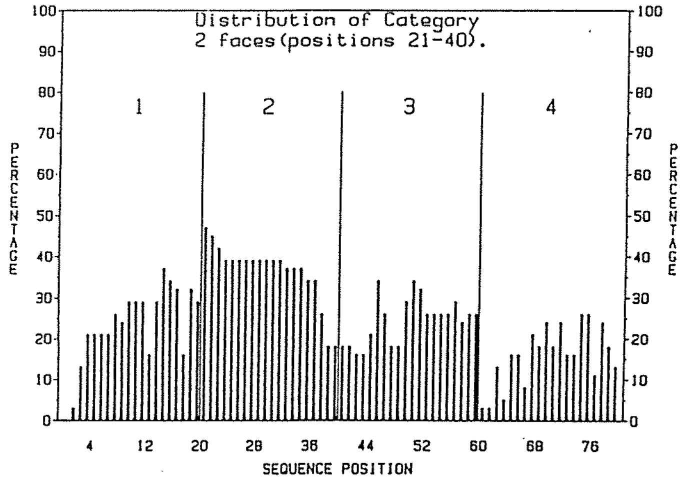
The four bar charts in Figure 1 show the percentage of subjects who agreed that each face belonged in a given category. This reflects the faces' "categorisability" - the degree to which subjects were able to discern those faces which were very similar to the anchor face, in contrast with those which were very dissimilar, or at some point in between. The "sequence position" in the diagrams is the final sequence of faces (1 - 80) arrived at by averaging the rankings of all subjects.

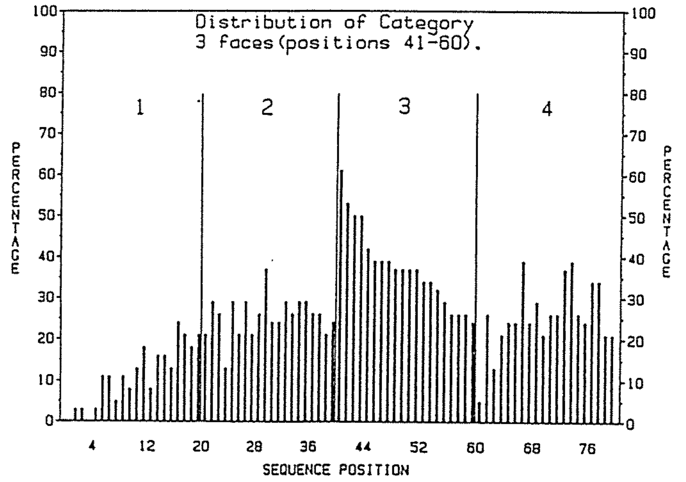
Each of the four diagrams shows the percentage of subjects who decided that a face in that final sequence position belonged in the category shown in that particular diagram. They also show the percentage of subjects who thought a final sequence face belonged in a different category. The divisions within each diagram show where a new category begins. It should be noted that the first face in the sequence was the anchor face provided to all subjects in position 1, which accounts for its lack of distribution over other categories.

Figure 1

The percentage of subjects agreeing that each face belongs in one of the four categories. The uppermost panel shows the percentage distribution for faces 1 - 20 (category 1) across the 80 possible sequence positions. The x-axis of the overall distribution is divided into four equal parts for convenience. It can be seen that most category 1 faces were placed in that category (rather than categories 2, 3, or 4) by a large majority of subjects. The other panels show similar data for faces that were finally placed in categories 2, 3, and 4. Categories 1 and 2 represent the high similarity sets, and categories 1 and 4 the low similarity sets. It is obvious from inspection that the distributions for categories 1 and 2 overlap to a greater extent than categories 1 and 4, suggesting a greater degree of separation, in terms of similarity judgments, between categories 1 and 4.


Distribution of Category
1 faces(positions 1-20).
1
2
3
4
PERCENTAGE
0
10
20
30
40
50
60
70
80
90
100
4
12
20
28
36
44
52
60
68
76
SEQUENCE POSITION


Distribution of Category
2 faces(positions 21-40).
1
2
3
4
PERCENTAGE
0
10
20
30
40
50
60
70
80
90
100
4
12
20
28
36
44
52
60
68
76
SEQUENCE POSITION


Distribution of Category
3 faces(positions 41-60).
1
2
3
4
PERCENTAGE
0
10
20
30
40
50
60
70
80
90
100
4
12
20
28
36
44
52
60
68
76
SEQUENCE POSITION

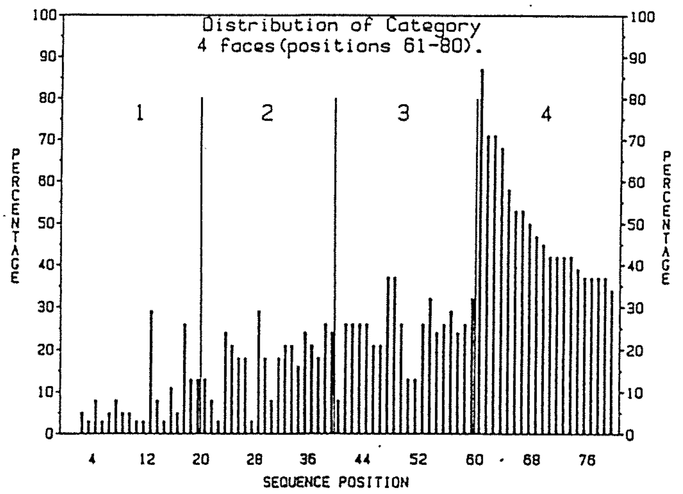

Distribution of Category
4 faces(positions 61-80).
1
2
3
4
PERCENTAGE
0
10
20
30
40
50
60
70
80
90
100
4
12
20
28
36
44
52
60
68
76
SEQUENCE POSITION

Figure 1 shows that a higher proportion of subjects were able to decide which faces were most and least similar to the anchor face, than were able to decide on intervening levels of similarity.

The uppermost distribution of Figure 1 shows for all the 80 faces of the final sequence, the proportion of subjects who, in their individual ordering of the stimuli, put faces into one of the first 20 positions, that is, selected them as being highest in similarity to the starting face. It can be seen that there were a few faces (e.g., those in final sequence positions 28, 39, 79, and 80) ending up in categories 2, 3, and 4, which a comparatively high proportion of subjects placed in one of the first 20 positions in their individual rankings. However, a mean of approximately 58% (from a high of 100% down to 37%) of subjects actually put the final sequence first 20 faces into the first 20 positions of their individual sequences.

The lowermost distribution can be interpreted similarly. The faces in final sequence positions 48 and 49 were selected by a comparatively high number of subjects to belong in the last 20 positions of their individual sequences, yet an average of approximately 50% (from 87% to 34%) of subjects put the last 20 faces of the final sequence into the last 20 positions of their own sequences.

The 40 faces occurring in between, those constituting categories 2 and 3, are not as well defined as categories 1 and 4. This makes sense when one considers that it is easier to rank stimuli at extremes rather than at intermediate levels. The faces in category 2, shown in the second panel of Figure 1, were selected by an average of approximately 36% (from 47% to 18%) of subjects. It can be seen that a number of faces in others of the final categories were also selected by equivalent proportions of subjects for category 2.

Category 3, shown in the third panel of Figure 1, came out little better. An average of about 37% (from 61% to 24%) of subjects selected category 3 faces in their own rankings. Again it can be seen that a number of faces in sequence positions outside category 3 were selected for that category by a comparatively large number of subjects.

In spite of the relatively poor definition of categories 2 and 3, separation did occur, which can clearly be seen comparing the distributions for categories 1 and 4. Overall, in only 9 out of the 80 faces did less than one-third of subjects select that face for its final category. Furthermore, for 24 faces, 50% or more of subjects selected those faces for their final categories. On the whole, this method does appear to be a reliable means of differentiating the stimuli in terms of similarity.

It was important that the degree of overlap between the target and two distractor categories should be quantified in some way. Ideally category 1 and 2 faces (high similarity) should be completely overlapping (so that they are indistinguishable) and categories 1 and 4 (low similarity) should not overlap at all.

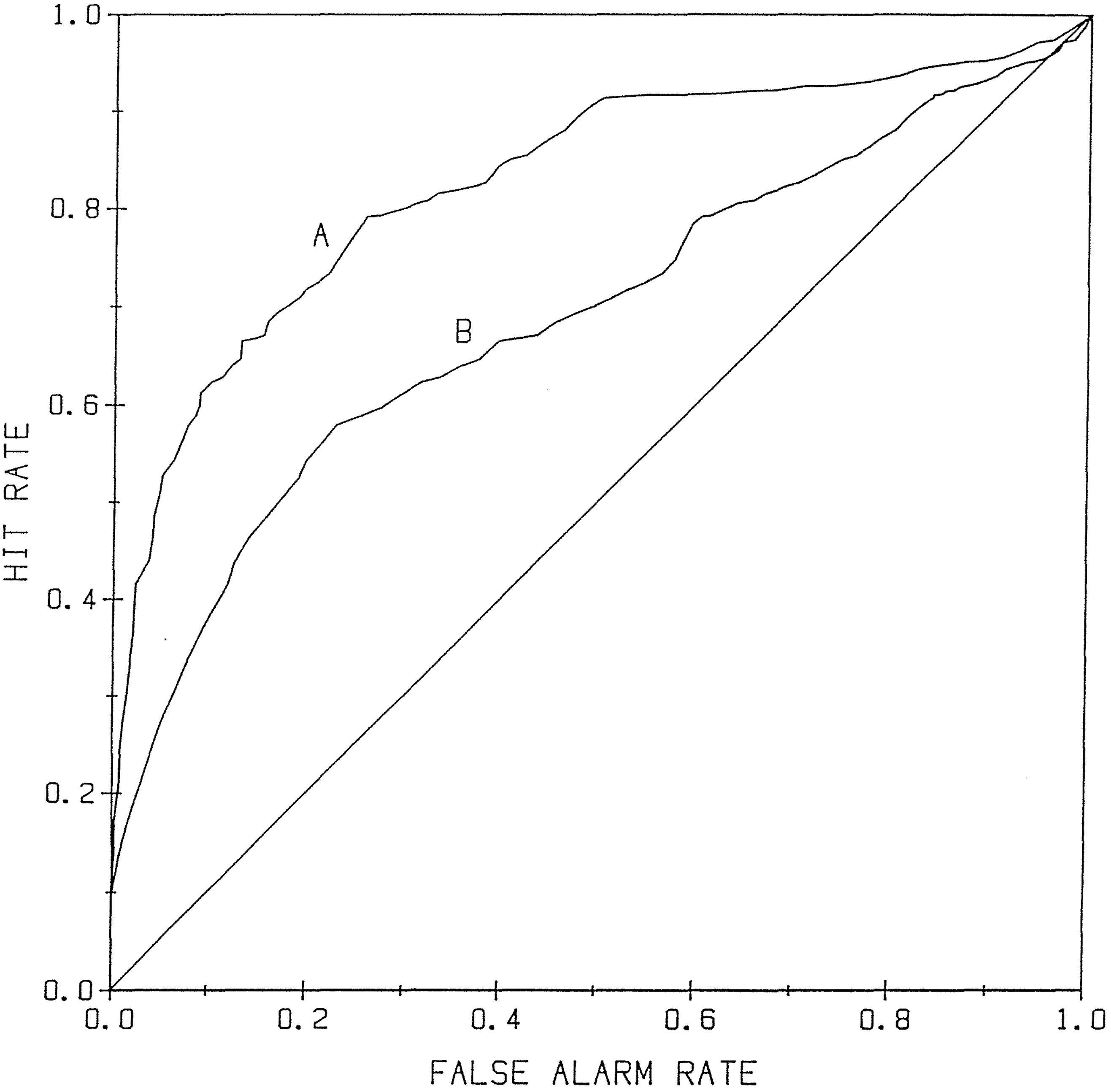
One way of measuring the overlap is to treat categories 1 and 2, and 1 and 4 as overlapping "signal" and "noise" distributions, as in Signal Detection Theory.

For example, categories 1 and 4 faces can be thought of as being distributed along a decision axis based on the similarity dimension. For the low similarity target / decoy sets the ideal would be to have perfect separation, in terms of similarity, between category 1 and category 4 faces. It can be seen from Figure 1 that this is not the case. However, the degree of overlap between the two distributions can be quantified by passing a criterion through the distributions (e.g., see McNicol, 1972) to generate an ROC curve.

The ROC curve is a plot of the cumulative hit rate against the cumulative false alarm rate. In the present case, a hit is defined (for a given criterion), as saying that a face was sampled from the category 1 distribution when in fact it was. A false alarm is defined as saying that a face was sampled from the category 1 distribution when in fact it was sampled from the category 2 (high similarity) or category 4 (low similarity) distribution.

Figure 2

ROC curves generated from pairs of overlapping distributions associated with the four categories of faces. For curve A, the hit rate represents the probability of saying a face was sampled from the distribution of category 1 faces, when in fact it was. The false alarm rate represents the probability of saying a face was sampled from the distribution of category 1 faces, when in fact the face belongs to the category 4 distribution. Similar definitions hold for curve B, based on the distributions of category 1 and category 2 faces. The area under the curves represents the degree of discriminability between the sets of faces. The area under curve A is 82%, and the area under curve B is 68%. See text for further details.


A
B
HIT RATE
FALSE ALARM RATE
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

The ROC curves generated by passing a criterion through the overlapping distributions based on categories 1 and 4 (curve A) and categories 1 and 2 (curve B) are shown in Figure 2. Had the distributions been non-overlapping (so that perfect discrimination between category 1 and 4 faces was possible) then the ROC curve would have traced the perimeter of the ROC space (0.0, 0.0; 0.0 1.0; 1.0, 1.0). Completely overlapping distributions would yield only chance performance and the ROC curve would follow the chance line in Figure 2 (0.0, 0.0; 1.0, 1.0). Ideally, category 1 and category 2 faces should not have been separable in terms of similarity.

As Figure 2 shows, both curves fail to meet the ideals described above. However, the area under curve A is a respectable 82%, and under curve B, 68%. The difference between the areas under the two curves reflects the degree to which the low similarity faces are more discriminable than the high similarity faces. The difference is 14%. Thus, it can be seen that the separation between the high and low target / decoy sets was less than ideal. Nevertheless, the results of Experiment II show that the difference obtained was sufficient to produce an effect on recognition rates for the high and low target / decoy sets.

So overall, the method, despite a lot of noise in the rankings, did yield noticeable differences between categories in terms of similarity.

The method used for ranking similarity had both virtues and weaknesses. The greatest virtue of the method was its ability to separate out faces of low similarity from those of high similarity to a given reference face. Only a small number of subjects put faces which the majority thought belonged in category 1 into category 4, or vice versa. As was required for Experiment II, a reasonably clear differentiation also emerged between categories 2 and 4. (Separation between categories 1 and 2 was not really an issue; it did not matter how much these overlapped. In fact, complete overlapping distributions would have yielded the maximum possible degree of similarity between target set and high similarity decoys.)

A major weakness of the ranking procedure was that it virtually gave subjects a free choice as to which of many dimensions they could use to judge similarity. This would have induced variability between subjects. An attempt was made to minimise this by suggesting three dimensions for subjects to use. Future research needs to investigate the utility of this. Also, this ranking procedure needs comparing with other methods, such as the method of triadic combinations, to further assess its reliability and validity.

A strong point of the method of analysis was its novel usage of SDT techniques for assessing separation between category 1 and 4 faces, and category 1 and 2 faces. ROC analysis seems ideal for quantifying overlap between distributions. The distributions in Figure 1 show how each category of faces (20 positions in the overall sequence of 80) were distributed in terms of choice across all 80 possible positions. While categories 1 and 4 were nicely separated, categories 2 and 4 were not so free of overlap.

It must be borne in mind, however, that there will always be some level of disagreement among subjects in any rating task based on a subjective dimension such as similarity. This was a very difficult task, given the extremely wide range of possible dimensions for basing decisions on. At chance level, each of the faces would have been allocated by subjects to each of the four categories 25% of the time. But as shown by the category groupings, the vast majority of faces in each of the categories were selected for that category by substantially more than 25% of subjects.

In conclusion, the main aim of Experiment I was to obtain similarity ratings on 80 faces relative to a starting face. This enabled us to categorize the faces into four groups and, as one might expect, the distribution and ROC curve analyses showed that while distinct groupings did emerge, the classification of many faces was prone to wide variability across subjects.

EXPERIMENT II

RELATIONSHIP BETWEEN SIMILARITY AND DELAY

As will be recalled, a major aim of this study was to test the theory that similarity may influence the effects of delay. Specifically, when target / decoy sets are dissimilar, they are more likely to decrease the impact of delay relative to target / decoy sets of higher similarity. The major impact is predicted to be upon the false alarm rate. Previous research suggests that with a range of delay periods and various numbers of targets and decoys, hit rate remains fairly stable. If recognition accuracy falls off it is mainly due to false alarm rate (Podd, 1990). Likewise for similarity; the theory predicts that target / decoy sets of high similarity will produce a greater false alarm rate than low similarity target / decoy sets.

Several clear predictions can be made from the theory.
(i) Delay will cause a rise in false alarm rate, leaving hit rate relatively constant.
(ii) Similarity also will produce a main effect for false alarm rate, but not for hit rate.
(iii) The combined effects of delay and target / decoy similarity should add together such that the false alarm rate for the longest delay period coupled with high target / decoy similarity should be greater than that for the same delay period coupled with low target / decoy similarity.

(iv) The clearest evidence in support of the theory would be an interaction between delay and similarity, such that delay produced a decrement in recognition rate, but only for the high similarity target / decoy sets.

It is difficult to predict clearly whether recognition accuracy will change. This is because a change in false alarm rate with no change in hit rate may not be sufficient to produce a significant change in A_g. In particular, low similarity target / decoy sets may produce no effects of delay over three weeks. However, on the basis of previous research (e.g., Podd, 1990; Davies et al., 1979), the expectation is that recognition accuracy will decrease with delay and with similarity. A clear prediction is that the highest false alarm rate will be for 21 day delay with high similarity target / decoy sets. The lowest false alarm rate will be for zero delay with low similarity target / decoy sets. A one day retention interval was also included in the study so as to examine the effect of a very short delay on similarity.

METHOD

Subjects

Ninety subjects took part in the second experiment, 15 per condition. Fifty-nine subjects were female, with 31 males; in each condition approximately one-third of subjects were male. Ages ranged from 17 to 44, with a mean of 22.1 years. All subjects were undergraduates at Massey University, the majority being students of psychology. All were volunteers, recruited by the experimenter from classes.

Materials and Design

The faces were presented in the form of slides by a Kodak Carousel projector onto a screen measuring 64 x 99 cm, which was 2.6m distant from the projector. Subjects were seated at tables between 2m and 3.2m distance from the screen, on which each face appeared approximately twice life-size. The projector was connected to an automatic timer, which presented each face singly, centre-screen, for a duration of five seconds, with a three second interval between trials.

A total of 60 slides were used; the target set of 20, the high similarity distractor set of 20, and the low similarity distractor set also of 20 faces. The 20 faces falling in category 3 were not used in the main study; their function was to serve to separate the two disparate groups of decoy faces (categories 2 and 4). Subjects recorded their responses on a two-page response sheet, numbered for trials 1 to 40, with four response categories for every trial:- "VERY SURE OLD" (1), "FAIRLY SURE OLD" (2), "FAIRLY SURE NEW" (3), AND "VERY SURE NEW" (4). (See Appendix A for sample response sheet.)

The independent variables were:

(1) The length of delay between study and test phases:- either zero, one, or 21 days; and

(2) The use of decoys that were either high or low in similarity to the target faces.

The dependent variables were the scores on the recognition test, represented by hit-rate, false alarm rate, and the recognition measures, d' and A_g. The experimental design was a 2 (High / Low Similarity) x 3 (0, 1, 21 days Delay) factorial design with both factors treated as between-subjects variables.

Procedure

There were two main phases in the experiment: a study phase, followed by a recognition phase.

The Study phase was the same for all subjects. In groups ranging in size from one to six members (all those within a group being from the same condition), they were shown the target set of 20 faces in a consecutive random order which remained constant for all subjects across all conditions.

As subjects arrived they were seated at a table in front of the screen. All subjects were aware that this was a study on facial recognition. After ensuring that all had a clear view of the screen, and answering any questions, the experimenter instructed subjects to "Please watch the screen carefully and try to remember the faces you see, as you will be asked to recognize them later on."

The lights were dimmed, and the projector and timer switched on. Each slide was projected onto the screen for five seconds, with a three second blank interval before the next slide. After showing all the target set, subjects were questioned as to whether they knew any of the people in the slides. Those who did were withdrawn and thanked for their participation in the study, with the explanation that they could not continue, as their familiarity with some faces, but not others, would result in a distorted recognition score. Subjects in other than the zero delay condition were thanked for attending this short session and asked to remember to return for the recognition phase either the next day or three weeks later, depending on which condition they were in.

Phase two, the Recognition phase, differed for subjects according to experimental condition. For those in the zero delay condition this took place immediately after phase one (with a five minute break in between). Subjects in the one day delay returned after 24 hours for the recognition phase; while 21 day delay subjects returned exactly three weeks later. In the recognition phase, half of the subjects (15) in each delay period saw high similarity distractors intermingled with the target faces, while the other half saw low similarity distractors with the targets. All subjects saw all the original 20 faces again, this time intermingled with decoy faces from either the high or the low similarity sets. The slides were shown in a previously determined random order (with the restriction that no more than three old faces or three new faces may occur consecutively) which was held constant across all delay conditions.

On their return, subjects were seated in the same position as for the study phase. This time on the table before them were a pen, a ruler, and a response form covered by the instruction sheet (see Appendix A for samples of these latter two). The experimenter read out the instructions, and then asked for any questions. The lights were once again dimmed (there was still sufficient light so the subjects could see to write their responses) and the timer and projector switched on for five blank trials. These blank trials were to give subjects an indication of how much time was available to look at each face slide and make a response to it. This was followed by a pause during which the experimenter checked that all subjects were sure of what was required of them, then the 40 target and distractor slides were shown.

At the conclusion, subjects were requested to write their age and sex on the response sheet. It was again ensured that none of the faces were familiar to any of them. At this stage, the subjects were debriefed, and informed of the hypotheses and expected results of the study. They were told how they could gain access both to their own personal score, and the results of the study as a whole. Any questions were answered, and finally the subjects were thanked for their participation.

RESULTS

Subjects' ratings on the response sheets were collapsed in order to calculate an overall hit rate and false alarm rate for each subject. Columns 1 and 2 were collapsed to yield "Old" responses, and columns 3 and 4 were collapsed to yield "New" responses (see Appendix A for response sheet). Hit rate is defined in this case as the subject choosing an "Old" response ("very sure old" and "fairly sure old"), given that the stimulus was a target. False alarm rate is defined as the subject responding "very sure old" or "fairly sure old" when the face was a decoy.

Hit rate and false alarm rate were also calculated for each rating category for each subject. The resulting pairs of hit and false alarm rates were used to plot 3-point ROC curves. Area under the ROC curve (A_g) is a relatively pure measure of recognition accuracy, being free of the potentially confounding effects of response bias.

The original intention of calculating d' from the overall hit rate and false alarm rate had to be abandoned because an unexpectedly large proportion of subjects performed at such a high level, scoring either 100% hits, 0% false alarms, or both. A 100% hit rate or 0% false alarm rate means that d' cannot be assessed without the use of some rather arbitrary "correction procedure". Thus, the A_g measure was the statistic used to assess recognizability.

Individual hit rate, false alarm rate, and A_g were each averaged across all subjects in each of the six conditions. These means and accompanying standard deviations are shown in Table 1. (Full individual data can be found in Appendix B.)

Table 1

Means and standard deviations for hit rate, false alarm rate, and A_g as a function of delay (in days) and target / distractor (T/D) similarity. An Arcsin transformation on A_g is also included. For each condition the upper figure is the mean, and the lower is the standard deviation.

	DELAY					
	0		1		21	
T/D Similarity	High	Low	High	Low	High	Low
Hits	.867	.867	.803	.890	.823	.847
	.090	.123	.137	.069	.098	.083
False Alarms	.110	.047	.127	.067	.120	.093
	.076	.048	.108	.067	.094	.053
A_g	.918	.957	.883	.956	.907	.917
	.057	.046	.116	.039	.048	.056
A_g(Arcsin)	2.590	2.787	2.510	2.779	2.537	2.586
	.204	.239	.352	.231	.157	.205

The theory that the degree of target / decoy similarity affects the level of recognition accuracy leads to some specific predictions for hit rate and false alarm rate.

Hit Rate

On the basis of previous findings and theoretical considerations the following was expected:

(a) No change in hit rate with high and low target / decoy similarity.

(b) No change in hit rate with delay.

Table 1 shows that no change occurred in hit rate at 0 delay for similarity. There was little change at 21 days, but nearly a 9% difference at 1 day. This was due to the fact that three subjects in the high similarity/1 day delay group scored 70% or less - an unusual occurrence within the same subject group. In fact, one subject scored only 45%. (If this subject is removed, the mean for that group increases to .829.) A two-way ANOVA was performed on hits (see Table 2).

Table 2

Summary table of two-way ANOVA for hits

Source	df	SS	MS	F	F-Prob
A Similarity	1	.03025	.03025	2.866	.094
B Delay	2	.01539	.00769	.729	.485
AB	2	.03017	.01508	1.429	.245
Sampling Error	84	.88667	.01056		
Total	89	.96247	.01081		

It can be seen from Table 1 that hit rate dropped by only 2% for low similarity and about 4% for high similarity groups across the 21 day delay. So, delay appeared not to markedly affect hit rate. This was confirmed by the ANOVAs, F(2,84) = 0.729, p = .485 for delay, and F(1,84) = 2.866, p = .094 for similarity. (See Table 2 for summary ANOVA table for hit rate.)

Thus, as predicted, there was no significant change in hit rate for either similarity or delay.

False Alarm Rate

The false alarm rate is the measure of greatest interest for the theory. It was predicted that:

(1) Delay would bring about an increase in false alarms.

(2) The high similarity target / decoy sets would produce a greater false alarm rate than the low similarity sets.

Thus, the highest false alarm rate in the study should have been for the high similarity 21 day delay, and lowest for the low similarity zero day delay.

Ignoring the 1 day delay for the moment, Table 1 shows that the main prediction was borne out. The lowest false alarm rate of 4.7% was for low similarity at zero delay, while the highest rate of 12% was for high similarity at 21 days delay. Collapsing over delay, the high similarity faces yielded an overall false alarm rate of 11.9%, while the low similarity faces produced 6.9%. Hence, there was a 5% increase in false alarms. An ANOVA on false alarms (see Table 3) showed this to be significant; F(1,84) = 9.369, p = .003.

Table 3

Summary table of two-way ANOVA for false alarms

Source	df	SS	MS	F	F-Prob
A Similarity	1	.05625	.05625	9.369	.003
B Delay	2	.01239	.00619	1.032	.361
AB	2	.00617	.00308	.514	.600
Sampling Error	84	.50433	.00600		
Total	89	.57914	.00651		

However, the effect of delay on false alarm rate, collapsed over similarity levels, was to increase false alarm rate from 7.9% at 0 delay, to 10.7% at 21 days delay. This was an increase of only 2.8%. This effect of delay on false alarm rate was not significant, F(2,84) = 1.032, p = .361. There was no interaction between similarity and delay (F<1). Thus, the major component in increasing false alarm rate was the effect of similarity, a 3 week delay adding little.

It is interesting to note that similarity appeared to have its greatest effects at the shorter delays of 0 and 1 day (see Table 1). However, the highest false alarm rate was obtained at 1 day delay for the high similarity target / decoy set. (Again, two abberrant false alarm scores of .40 and .25, the highest and second equal highest false alarm rate across all subjects, boosted false alarm rate at 1 day delay.)

In summary, there was an increase in false alarm rate for high similarity faces as predicted, but, surprisingly, no increase in false alarms at 21 days delay.

Recognition Accuracy

Recognition accuracy was assessed by the measure A_g. The theory makes the prediction that both the effects of similarity and delay will decrease recognition accuracy, mainly due to a decrease in false alarm rate. With the possibility of having a significant change in false alarm rate and hit rate remaining constant, the recognition accuracy measure may or may not change significantly.

As Table 1 shows, collapsed over similarity, delay over 21 days decreased A_g by only 2.6% (0 delay, A_g = 93.8%; 21 days, A_g = 91.2%). The low 88.3% A_g score at 1 day delay, high similarity was due to an abberrant subject obtaining only a 52% A_g score. (With this subject's score removed, the mean for the 1 day group moves from 88.3% to 90.8%.) Table 4 presents the ANOVA for A_g.

Table 4

Summary table of two-way ANOVA for A_g

Source	df	SS	MS	F	F-Prob
A Similarity	1	.03711	.03711	8.676	.004
B Delay	2	.01040	.00520	1.215	.302
AB	2	.01512	.00756	1.767	.177
Sampling Error	84	.35930	.00428		
Total	89	.42193	.00474		

The ANOVA of delay on A_g confirms that there was no effect, F(2,84) = 1.215, p = .302. However, the average effect of similarity, collapsed over delay, was a 4% drop in A_g for the high similarity target / decoy set, F(1,84) = 8.676, p = .004.

Thus, as might have been expected from the analysis of hit and false alarm rates, high similarity target / decoy sets produced a small but significant decrease in recognition accuracy compared to low similarity sets. The 21 day delay failed to have any effect on recognition accuracy.

McNicol (1972) points to a problem which can arise when using a measure that is a probability score (thus having an upper limit of 1) as a sensitivity score, which is the case with the A_g measure. When some conditions yield high sensitivity, the distribution of subjects' scores will be positively skewed. McNicol states that this skewness "can have unfortunate effects on the subsequent analysis of variance, so that it is best to remove it first." (p.118).

The usual way of doing this is to transform the raw scores using 2 arcsin $\sqrt{p}$, where p is the raw probability. The arcsin transformation has an upper limit of π, rather than 1, so it stretches out the distribution at the top end. Because the recognition accuracy scores were mostly very high, it was thought prudent to carry out the arcsin transformation to see if it had any effect on the ANOVA. Table 5 shows the results of an ANOVA conducted on the transformed scores.

Table 5

Summary table of two-way ANOVA for arcsin transformation of A_g.

Source	df	SS	MS	F	F-Prob
A Similarity	1	.65938	.65938	11.562	.001
B Delay	2	.25083	.12542	2.199	.117
AB	2	.18750	.09375	1.644	.199
Sampling Error	84	4.79064	.05703		
Total	89	5.88835	.06616		

It can be seen by comparing Tables 4 and 5 that the transformation made no real difference, except to slightly increase the F values for similarity and delay.

Finally, a check was made to find out if any condition caused subjects to be biased toward reporting faces as "Old" or "New". Because subjects used a 4-point rating scale, it was not possible to use a conventional bias estimate such as Beta. One simple method is to calculate the frequency of "Old" and "New" responses. Since old and new faces were presented with equal probability and the cost of making a "New" response was no different from an "Old" response, 50% "Old" and 50% "New" responses were expected.

Table 6

Means and standard deviations for % "New" responses as a function of delay (in days) and target / distractor (T/D) similarity. For each condition the upper figure is the mean, and the lower is the standard deviation.

	DELAY					
	0		1		21	
T/D Similarity	High	Low	High	Low	High	Low
	.512	.543	.542	.522	.528	.532
	.053	.064	.067	.042	.076	.042

Table 6 shows no apparent effect across similarity or delay, but it should be noted that there is a slight bias toward reporting a face as "New" across all conditions. A similar small but consistent bias to reporting faces as "New" was noted by LaMontagne (1989) and Podd (1990).

Table 7 shows the results of the ANOVA which confirmed that neither similarity nor delay affected % "New" (in both cases F < 1). There was no interaction.

Table 7

Summary table of two-way ANOVA for % "New" responses

Source	df	SS	MS	F	F-Prob
A Similarity	1	5.62500	5.62500	.163	.688
B Delay	2	2.63889	1.31944	.038	.963
AB	2	100.41667	50.20833	1.453	.240
Sampling Error	84	2902.50000	34.55357		
Total	89	3011.18056	33.83349		

DISCUSSION

Hit Rate

As predicted by the theory, neither delay nor similarity had any effect on hit rate. Under the theory, subjects match previously seen faces to their images stored in memory. Because the images of the target faces are well retained in memory over a relatively short delay, hit rate is unimpaired. The presence of high or low similarity decoys does not affect subjects' ability to identify targets, so likewise has no effect on hit rate.

The results of the present study and the findings of previous researchers appear to bear out this theory. Egan et al. (1977), Chance and Goldstein (1987), and Podd (1990) found that hit rate remained relatively constant with delay, while Davies et al. (1979) found no effect for similarity on hit rate.

False Alarm Rate

Similarity had the predicted effect on false alarms. High similarity produced a greater false alarm rate with the greater likelihood that a highly similar decoy would match the memory trace of the target face. Egan et al. (1977), Cohen and Nodine (1978), Davies et al. (1979), and Light et al. (1979) also found that high similarity led to an increase in false alarms.

The major results with respect to delay were very surprising. There were no significant effects after three weeks for either high or low similarity faces, whereas previous research on delay has shown effects for a retention interval in the region of three weeks. Davies et al. (1978), Courtois and Mueller (1981), and Podd (1990), all obtained an effect for delay over intervals of 21, 28, and 14 days respectively.

LaMontagne (1989) used 70 photographs drawn from the same photo pool (100 faces) as the present study and got an effect for 35 targets / 35 distractors over three weeks. However, he used black and white faces, whereas colour photographs were used in the present experiment, and he used larger target / distractor sets. The additional dimensions of hair colour, eye colour, and skin complexion may have provided further memory cues to help subjects perform well, even over the comparatively long three week delay. In effect, the task was too easy, with the presence of colour possibly increasing the effective level of dissimilarity between faces by increasing the number of cues available for making the discrimination.

The lack of effect of delay for low similarity faces could have been borne, but the absence of an effect for high similarity faces (i.e., an interaction between delay and similarity) does not enable a proper test of the theory. The false alarm results are equivocal regarding supporting or refuting the theory. The theory predicted that the greater the degree of similarity between target and distractor sets, the more a distractor is likely to be called a target, thus increasing false alarms, but leaving hit rate unaffected. Also predicted was a greater increase in false alarms than hits for the effects of delay. Building on this, long delays and high similarity should have an additive effect, yielding the highest false alarm rate.

From the results we can see that indeed false alarms were more affected by similarity than was hit rate. Unfortunately, as delay produced no significant effects, and there was no significant interaction between delay and similarity, the second part of the theory could not be adequately tested. It therefore follows that we could not evaluate the prediction that long delays and high similarity would show an additive effect, such that the effects of delay would be more pronounced for high similarity target / decoy sets, because delay appeared to produce no effect for either level of similarity. The data are suggestive of these effects, and it would be wrong to dismiss them merely because the results are not significant at the .05 level. The correct procedure is to withhold a conclusion until we have more evidence (Hays, 1973). Further work is required, perhaps using longer delay periods, to more adequately test the theory.

Area Under the ROC Curve

Although the theory made no definite predictions for recognition accuracy, similarity did have the expected effect by producing a significant drop in A_g. It is important to note here that the major effect on the recognizability index was the increased false alarm rate. That is, recognizability dropped with high similarity target / decoy sets because subjects called more decoys targets, and not because they could not recall the targets themselves. This suggests that increasing the degree of target / decoy similarity has little effect on the encoding, storage, and retrieval of targets. However, it seems that interference effects do occur at the retrieval stage with high similarity decoys.

It was surprising that no such decrease occurred with delay. This result drives home the fact that subjects appear to have a remarkable recognition memory for faces.

Although some of the results showing the effect of delay were in the expected direction, they were not quite strong enough to "reach significance." Thus any further interpretation of the findings relating to similarity could not be carried out because they built on the effect of delay, which, if the .05 rule is adhered to, was a non-effect.

One has to wonder whether a rigid adherence to the .05 rule is justified when examining behaviour changes as a function of time. For example, over a delay of a few days it may be that only very small changes occur in the measures. They may not be sufficient to be "significant", but if they are consistent they can be considered to be real. It might be more useful for future research to determine a theoretical function for recognition accuracy as a function of time. Data could be tested by goodness of fit measures, rather than whether a change from time 1 to time 2 is significant. It may well be statistically significant if the forgetting function is very steep between time 1 and time 2, but not if the function is shallow.

In summary, the results of Experiment II offer mixed support for the theory. In terms of hit rate, the findings were much as expected. The results pertaining to false alarm rate were as predicted for similarity, but there were no effects for delay. To adequately test the theory, it was necessary to obtain an effect on false alarm rate for either low or high similarity, or both. It could well be that the effect of delay on false alarm rate varies across different delay periods. It is necessary to determine exactly which delay periods produce the largest effects on false alarm rate. It is also possible that even the "high" similarity set was not in fact very similar to the target set at all, which may explain why no effect was obtained for delay. The only way of finding out is to make categories 1 and 2 more similar.

Suggestions for Further Work

As outlined above, it was not possible to properly test the theory because of the absence of an effect for delay. Ways of preventing this in future research shall now be considered.

The research currently in existence involves a great deal of uncertainty. Apart from a lack of control over third variables (e.g., similarity between faces) present in many studies, it remains unclear exactly what are the appropriate delay periods to use. For instance Yarmey (1979) considers one week to be tapping long term memory, and one month as very long term. If so, one must ask what were Bahrick, Bahrick, and Wittlinger (1975) investigating in their study "Fifty Years of Memory for Names and Faces"? And Davies, Ellis, and Shepherd (1978) consider anything up to two days to be "immediate"!

Needed is some way of consolidating the plethora of data which documents the decline in recognition performance over time. Procedural differences between studies make comparisons difficult. The major question which needs answering with regard to retention interval for faces, and which has so far been largely unaddressed, is this:- Does the rate of forgetting for faces follow the standard Ebbinghaus forgetting curve? Such a basic question needs to be answered before research can meaningfully investigate problems involving delay and facial recognition. What is required is research using a large range of delay intervals from short to long term, so that the rate of forgetting for faces may be accurately plotted.

Specifically, such a study should have periods of delay stretching at weekly intervals over a long period (ten weeks or so). It would be interesting to examine the functions for hit rate, false alarm rate, and recognition accuracy over this period. Even though Podd's (1990) research suggests that short delays may first affect false alarm rate, a point must come where hit rate starts to decline also. It is also important to find out just how target / decoy similarity interacts with delay. The current theory suggests that where target / decoy similarity is low, then one may not obtain an effect for delay. Research over extended retention intervals is needed to properly test this, and to find out over what range of delays (if any) the theory applies.

The scaling of similarity in the present study seemed successful. However, in hindsight, it may have been beneficial to interleave categories 1 and 2, inserting category 2 faces in between category 1 faces. Then the first twenty faces, which are the targets, would be composed equally of category 1 and category 2 faces, and likewise for the second twenty faces, the high similarity decoys. This overlapping of categories would have had the effect of making these two sets maximally similar to each other, so producing, hopefully, a stronger similarity effect, which, under the theory, ought to have increased the effect of delay.

A major weakness of the present study was the ceiling effects produced by unexpectedly high recognition accuracy. This can probably be partly explained by the use of colour photographs. Further research needs to look into colour / black and white differences.

With longer delays and/or shorter stimulus presentation times, it should be possible to reduce the very high recognition rates which unexpectedly occurred in the present study, thus preventing ceiling effects. The theory can be properly evaluated only when delay produces an effect in either low or high similarity target / decoy groups, or in both.

Thus to give the theory a more adequate test, and to look at hit rate and false alarm rate over time, the following study is suggested. Interleave category 1 and 2 faces to produce stronger differences between high and low similarity. Face

stimuli should be presented for shorter intervals than in the present study (for instance, show the faces for only two seconds each, instead of five seconds) to obtain lower recognition rates. More retention intervals should be used, spanning a longer period of time - perhaps up to 10 weeks. Such a study would answer the following important questions:
(a) What are the functions of hit rate and false alarm rate over delays varying from zero to several weeks?
(b) What periods of delay are affected the most by high or low similarity target / decoy sets?
In addition, such a study would provide some basic parametric data on delay - of which there is a deficiency in the literature.

Summary and Conclusions

The present research aimed to find out why the effects of delay in facial recogniion research are inconsistent. A theory was developed which predicted that high similarity target / decoy sets may cause relatively high false alarm rates, and thus an effect for delay, whereas low similarity target / decoy sets generate low false alarm rates and may not produce an effect of delay. In other words, it was argued that the effect of delay (for some delay periods at least) may be influenced by the degree of target / decoy similarity.

A major exercise to scale 80 faces for similarity was conducted successfully before attempting to test the theory. While there were clear differences between the high/low similarity target / decoy ensembles, there was no main effect for delay on any of the measures, nor an interaction between similarity and delay. These unexpected results meant that the experiment could not adequately test the theory, since delay had no effect for either high or low similarity sets.

The following conclusions can be drawn from these two studies.
(a) It seems possible to successfully manipulate target / decoy similarity through a relatively simple rating procedure. This is extraordinary, given that subjects were not forced to use any set of criteria, although three were suggested. Further evidence that success occurred here was gained from Experiment II, where the different levels of target / decoy similarity determined by the scaling procedure did indeed produce the predicted effects.
(b) The study was less successful in testing the theory proposed. Results were excellent for similarity, but because no significant effects of delay occurred for either the high or the low similarity target / decoy sets, the theory could not be properly evaluated. The major reason is that the degree of similarity between targets and decoys may not have been sufficient.

Lack of time precluded the running of a more appropriate study, but one has been suggested. Finally, it is recommended that more emphasis should be placed on the shape of the forgetting curve, rather than whether "significant" differences occur between various points on the curve.

REFERENCES

Bahrick, H.P., Bahrick, P.O., and Wittlinger, R.P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. Journal of Experimental Psychology : General, 104, 54-75.

Banks, W.P. (1970). Signal detection theory and human memory. Psychological Bulletin, 74, 81-99.

Barkowitz, P. and Brigham, J.C. (1982). Recognition of faces: Own-race bias, incentive, and time delay. Journal of Applied Social Psychology, 12, 255-268.

Buckhout, R., Alper, A., Chern, S., Silverberg, G., and Slomovits, M. (1974). Determinants of eyewitness performance on a lineup. Bulletin of the Psychonomic Society, 4, 191-192.

Chance, J.E., and Goldstein, A.G. (1987). Retention interval and face recognition: Response latency measures. Bulletin of the Psychonomic Society, 25, 415-418.

Chance, J., Goldstein, A.G., and McBride, L. (1975). Differential experience and recognition memory for faces. Journal of Social Psychology, 97, 243-253.

Cohen, M.E. and Nodine, C.F. (1978). Memory processes in facial recognition and recall. Bulletin of the Psychonomic Society, 12, 317-319.

Courtois, M.R. and Mueller, J.H. (1981). Target and distractor typicality in facial recognition. Journal of Applied Psychology, 66, 639-645.

Davies, G.M., Ellis, H.D., and Shepherd, J.W. (1978). Face identification: The influence of delay upon accuracy of Photofit construction. Journal of Police Science and Administration, 6, 35-42.

Davies, G.M., Ellis, H.D., and Shepherd, J.W. (1981). Perceiving and remembering faces. London: Academic Press.

Davies, G.M., Shepherd, J.W., and Ellis, H.D. (1979). Similarity effects in facial recognition. American Journal of Psychology, 92, 507-523.

Deffenbacher, K.A. (1986). On the memorability of the human face. In H.D. Ellis, M.A. Jeeves, F. Newcombe, and A. Young (Eds.), Aspects of face processing. Dordrecht: Martinus Nijhoff.

Deffenbacher, K.A., Carr, T.H., and Leu, J.R. (1981). Memory for words, pictures, and faces: Retroactive interference, forgetting, and reminiscence. Journal of Experimental Psychology: Human Learning and Memory, 7, 299-305.

Egan, D., Pittner, M., and Goldstein, A.G. (1977). Eyewitness identification: Photographs vs. live models. Law and Human Behaviour, 1, 199-206.

Ellis, H.D. (1975). Recognizing faces. British Journal of Psychology, 66, 409-426.

Ellis, H.D. (1981). Introduction. In G. Davies, H. Ellis, and J. Shepherd (Eds.). Perceiving and remembering faces. London: Academic Press.

Ellis, H.D. (1984). Practical aspects of face memory. In G.L. Wells and E.F. Loftus (Eds.), Eyewitness testimony: psychological perspectives. New York: Cambridge University Press.

Ellis, H. D. (1986). Introduction to aspects of face processing. In H.D. Ellis, M.A. Jeeves, F. Newcombe, and A. Young (Eds.), Aspects of face processing. Dordrecht: Martinus Nijhoff.

Goldstein, A.G. and Chance, J.E. (1970). Visual recognition memory for complex configurations. Perception and Psychophysics, 9, 237-241.

Goldstein, A.G. and Chance, J.E. (1981). Laboratory studies of face recognition. In G.M. Davies, H.D. Ellis, and J.W. Shepherd (Eds.), Perceiving and remembering faces. London: Academic Press.

Goldstein, A.G., Stephenson, B., and Chance, J.E. (1977). Face recognition memory: Distribution of false alarms. Bulletin of the Psychonomic Society, 9, 416-418.

Green, D.M. and Swets, J.A. (1966). Signal detection theory and psychophysics· New York: Wiley.

Johnson, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 32, 241-254.

Krouse, F.L. (1981). Effects of pose, pose change, and delay on face recognition performance. Journal of Applied Psychology, 66, 651-654.

LaMontagne, H. (1989). The effect of Photofit-type faces on recognition memory. Unpublished Masters thesis, Massey University, Palmerston North.

Laughery, K.R., Fessler, P.K., Lenorovitz, D.R., and Yoblick, D.A. (1974). Time delay and similarity effects in face recognition. Journal of Applied Psychology, 59, 490-496.

Light, L.L., Kayra-Stuart, F., and Hollander, S. (1979). Recognition memory for typical and unusual faces. Journal of Experimental Psychology: Human Learning and Memory, 5, 212-228.

Lipton, J.P. (1977). On the psychology of eyewitness testimony. Journal of Applied Psychology, 62, 90-95.

McNicol, D. (1972). A primer of signal detection theory. London: Allen and Unwin.

Podd, J. (1990). The effects of memory load and delay on facial recognition. Journal of Applied Cognitive Psychology, 4, 47-59.

Read, J.D. (1979). Rehearsal and recognition of human faces. American Journal of Psychology, 92, 71-85.

Scapinello, K.F. and Yarmey, A.D. (1970). The role of familiarity and orientation in immediate and delayed recognition of pictorial stimuli. Psychonomic Science, 21, 329-330.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. British Journal of Psychology, 75, 221-242.

Shapiro, P.N. and Penrod, S. (1986). Meta-analysis of facial identification studies. Psychological Bulletin, 100, 139-156.

Shepherd, J.W. (1983). Identification after long delays. In S.M.A. Lloyd-Bostock and B.R. Clifford (Eds.), Evaluating witness evidence. Chichester: Wiley.

Shepherd, J., Davies, G., and Ellis, H. (1981). Studies of cue saliency. In G. Davies, H. Ellis, and J. Shepherd (Eds.). Perceiving and remembering faces. London: Academic Press.

Shepherd, J.W. and Ellis, H.D. (1973). The effect of attractiveness on recognition memory for faces. American Journal of Psychology, 86, 627-633.

Shepherd, J.W., Ellis, H.D., and Davies, G.M. (1982). Identification evidence: A psychological evaluation. Aberdeen: The University Press.

Thomson, D.M. (1986). Face recognition: More than a feeling of familiarity? In H.D. Ellis, M.A. Jeeves, F. Newcombe, and A. Young (Eds.), Aspects of face processing. Dordrecht: Martinus Nijhoff.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: Wiley.

Walker-Smith, G.J. (1978). The effects of delay and exposure duration in a face recognition task. Perception and Psychophysics, 24, 63-70.

Wallace, G., Coltheart, M., and Forster, K.I. (1970). Reminiscence in recognition memory for faces. Psychonomic Science, 18, 335-336.

Wells, G.L. and Loftus, C.F. (1984). Eyewitness testimony: psychological perspectives. New York: Cambridge University Press.

Yarmey, A.D. (1971). Recognition memory for familiar 'public' faces: Effects of orientation and delay. Psychonomic Science, 24, 286-288.

Yarmey, A.D. (1979a). The effects of attractiveness, feature saliency and liking on memory for faces. In M. Cook and G. Wilson (Eds.), Love and attraction. New York: Pergamon Press.

Yarmey, A.D. (1979b). The psychology of eyewitness testimony. New York: The Free Press.

APPENDICES

APPENDIX A

Instructions and Response Sheet

(Note that only the first page of the response sheet is shown.

The second page is identical apart from numbering.)

INSTRUCTIONS

THIS PART OF THE EXPERIMENT WILL TEST HOW WELL YOU CAN RECOGNIZE FACES YOU SAW IN THE FIRST PART OF THE EXPERIMENT. FACES YOU HAVE ALREADY SEEN ARE CALLED OLD FACES. THESE ARE MIXED IN WITH FACES YOU HAVE NOT PREVIOUSLY SEEN. THESE ARE CALLED NEW FACES.

YOUR TASK IS TO RATE HOW CERTAIN YOU ARE THAT EACH FACE IS OLD (PREVIOUSLY SEEN) OR NEW (NOT PREVIOUSLY SEEN). TRY TO USE ALL THE RATING CATEGORIES AVAILABLE - VERY SURE OLD (1), FAIRLY SURE OLD (2), FAIRLY SURE NEW (3), AND VERY SURE NEW (4).

ON EACH TRIAL, THE PRESENTATION OF AN OLD OR A NEW FACE IS EQUALLY LIKELY; THAT IS 50/50.

EACH FACE WILL BE PRESENTED FOR 5 SECONDS FOLLOWED BY A 3 SECOND DECISION INTERVAL DURING WHICH YOU CAN RECORD YOUR RESPONSE.

THE FIRST FIVE TRIALS ARE FOR PRACTICE PURPOSES ONLY. THESE TRIALS WITH BLANK SLIDES WILL GIVE YOU AN IDEA OF HOW LONG EACH SLIDE WILL BE PRESENTED AND THE TIME YOU HAVE TO RESPOND.

ANY QUESTIONS?

TRIAL.	VERY SURE OLD	FAIRLY SURE OLD	FAIRLY SURE NEW	VERY SURE NEW
1.	1	2	3	4
2.	1	2	3	4
3.	1	2	3	4
4.	1	2	3	4
5.	1	2	3	4
6.	1	2	3	4
7.	1	2	3	4
8.	1	2	3	4
9.	1	2	3	4
10.	1	2	3	4
11.	1	2	3	4
12.	1	2	3	4
13.	1	2	3	4
14.	1	2	3	4
15.	1	2	3	4
16.	1	2	3	4
17.	1	2	3	4
18.	1	2	3	4
19.	1	2	3	4
20.	1	2	3	4
	VERY SURE OLD	FAIRLY SURE OLD	FAIRLY SURE NEW	VERY SURE NEW

APPENDIX B

Individual Data for Subjects by Condition

Below are shown the full data for each subject (s) in each of the six treatment conditions. Given are their scores in terms of hit rate (HR), false alarm rate (FAR), area under the ROC curve (A_g), and percentage of "New" responses (New). Means and standard deviations for these measures are also included.

Low similarity - 0 delay condition

	HR	FAR	A_g	New
s1	1.000	.100	.988	.450
s2	.750	.000	.938	.625
s3	.900	.000	.990	.550
s4	.600	.050	.831	.675
s5	.950	.000	1.000	.525
s6	1.000	.100	.988	.450
s7	.800	.100	.903	.550
s8	.950	.000	.974	.525
s9	.950	.000	.998	.525
s10	.950	.100	.943	.475
s11	.850	.100	.945	.525
s12	1.000	.000	1.000	.500
s13	.850	.000	.978	.575
s14	.750	.100	.933	.575
s15	.700	.050	.944	.625
mean	.867	.047	.957	.543
s.d.	.123	.048	.046	.064

High similarity - 0 delay condition

	HR	FAR	A_g	New
s1	.850	.150	.913	.500
s2	.950	.200	.953	.425
s3	.950	.050	.968	.500
s4	.800	.200	.845	.500
s5	.850	.150	.875	.500
s6	.900	.150	.954	.475
s7	.850	.050	.950	.550
s8	.950	.000	.993	.525
s9	.900	.150	.943	.475
s10	.800	.100	.885	.550
s11	.900	.050	.943	.525
s12	.900	.000	.960	.550
s13	.950	.050	.936	.500
s14	.600	.100	.775	.650
s15	.850	.250	.878	.450
mean	.867	.110	.918	.512
s.d.	.089	.076	.057	.053

Low similarity - one day delay condition

	HR	FAR	A_g	New
s1	.950	.050	.995	.500
s2	.900	.100	.933	.500
s3	.850	.050	.956	.550
s4	.850	.150	.919	.500
s5	.800	.050	.944	.575
s6	.900	.150	.935	.475
s7	1.000	.000	1.000	.500
s8	.850	.050	.964	.550
s9	.800	.150	.883	.525
s10	.800	.050	.965	.575
s11	.900	.000	.993	.550
s12	1.000	.000	1.000	.500
s13	.950	.000	.998	.525
s14	.850	.000	.966	.575
s15	.950	.200	.888	.425
mean	.890	.067	.956	.522
s.d.	.069	.067	.039	.042

High similarity - one day delay condition

	HR	FAR	A_g	New
s1	.900	.000	.978	.550
s2	.850	.200	.931	.575
s3	.800	.050	.875	.575
s4	.750	.000	.906	.625
s5	1.000	.100	1.000	.450
s6	.650	.150	.790	.600
s7	.450	.400	.524	.575
s8	.750	.100	.895	.575
s9	.800	.100	.895	.550
s10	1.000	.200	.985	.400
s11	.850	.250	.888	.450
s12	.850	.050	.911	.550
s13	.850	.200	.894	.475
s14	.700	.050	.801	.625
s15	.850	.050	.965	.550
mean	.803	.127	.883	.542
s.d.	.136	.108	.116	.067

Low similarity - 21 day delay condition

	HR	FAR	A_g	New
s1	.950	.050	.984	.500
s2	.900	.000	.930	.550
s3	.700	.150	.778	.575
s4	.900	.200	.920	.475
s5	.800	.050	.928	.575
s6	.950	.100	.959	.475
s7	.850	.100	.911	.525
s8	.900	.100	.943	.500
s9	.900	.050	.954	.525
s10	.800	.050	.950	.575
s11	.900	.050	.991	.525
s12	.800	.150	.895	.525
s13	.750	.100	.883	.575
s14	.900	.150	.884	.475
s15	.700	.100	.840	.600
mean	.847	.093	.917	.532
s.d.	.083	.053	.056	.042

High similarity - 21 day delay condition

	HR	FAR	A_g	New
s1	.950	.050	.943	.500
s2	.900	.250	.906	.425
s3	.700	.100	.853	.600
s4	.800	.100	.873	.550
s5	.950	.150	.965	.450
s6	.800	.050	.894	.575
s7	.900	.150	.943	.475
s8	.850	.100	.909	.525
s9	.800	.000	.930	.600
s10	.900	.100	.940	.500
s11	.650	.250	.795	.550
s12	.700	.000	.928	.650
s13	.700	.000	.943	.650
s14	.850	.250	.839	.450
s15	.900	.250	.943	.425
mean	.823	.120	.907	.528
s.d.	.098	.094	.048	.076