Distinctiveness of a face predicts both miss errors (MEs) and false positives (FPs) but correlations between these errors are low (e.g. Hancock, Burton, & Bruce, 1996). To investigate this, distinctiveness and personal familiarity were analysed as predictors of MEs and FPs in a face recognition experiment. Faces were assigned to three groups, which meant that each set were distractor faces for two different sets of targets. Mean ratings of distinctiveness predicted MEs, whereas familiarity predicted FPs only if individual ratings were used. The degree to which subjects were consistent in their ratings and performance over different faces was also considered. Good subject consistency was found on FPs when the subjects saw the same target faces. If subjects who had seen different target faces were compared, then the consistency of FPs was lower than the consistency of MEs. The results imply that distinctiveness predicts MEs as a general property of the population of faces, whereas familiarity predicts FPs according to the idiosyncrasies of subjects.

INTRODUCTION

Mistaking an unknown person as someone familiar is a common occurrence. Indeed, Young, Hay and Ellis (1985) found that 30% of reported errors of recognition were of this type of false positive recognition (or...
false alarms). This phenomenon usually leads to little more than slight embarrassment; however, cases of false positive recognition can have much more serious consequences, such as miscarriages of justice. The present study investigated what factors predict production of false positive errors in face recognition with the aim of adding to our understanding of how face recognition takes place.

In many face recognition experiments, subjects have the opportunity to make false positive responses. A false positive is a response to a previously unseen distractor item as if it were a previously seen target item. Sometimes these errors are not analysed due to insufficient numbers, but some studies have looked at false positives (e.g. Bartlett & Fulton, 1991; Bartlett, Hurry, & Thorley, 1984; Goldstein & Chance, 1980; Hancock, Burton, & Bruce, 1996; Light, Kayra-Stuart, & Hollander, 1979; Vokey & Read, 1988, 1992; Yin, 1969). Overall, these studies found that there are at least two candidates which may be strong predictors of the production of false positives: the distinctiveness of the faces and their rated familiarity.

Before the roles of these factors are discussed, it would be useful to clarify what is meant by these and other terms used. A distinctiveness rating of a face is gained by asking subjects how easy it would be to spot that person in a crowd. A distinctive face would stand out from a crowd, whereas a typical face would be very similar to many other faces. Distinctiveness is an antonym for typicality used by Vokey and Read (1992).

The basic term “familiarity” can have more than one meaning. In its usual sense, it refers to a feeling of having seen something before. During a recognition experiment, this “perceived familiarity” or “specific familiarity” will lead to recognition of a target item. Another meaning of familiarity is “context-free familiarity”. Bartlett et al. (1984) defined this as familiarity to a previously unseen face and so it is the feeling of having possibly seen something before when actually it is a novel item. Vokey and Read (1992) use the term “general familiarity” to describe this rateable property of faces.

To illustrate the differences between the two forms of familiarity, it is useful to consider the rating procedures employed by Bartlett and Fulton (1991). In their experiments, subjects rated each face on whether it occurred earlier in the sequence (i.e. specific familiarity or recognition) and also whether the face was confusable with someone they knew (i.e. general recognition). This measure of general familiarity was referred to as “subjective familiarity” by Bartlett and Fulton.

For the purposes of this experiment, it is necessary to introduce a new form of familiarity called “personal familiarity”. This is the familiarity that a face has due to its similarity to a face known by the subject prior to starting the experiment. Since subjects will know different faces, then
this rating should be personal to that subject. To obtain ratings of personal familiarity, subjects are asked whether the face looks like someone they know. Within the experiment, this personal familiarity will not lead to recognition because the context is wrong for that face, and so this is similar to context-free familiarity as described by Bartlett et al. (1984). This means that personal familiarity is more closely related to general than specific familiarity, but it differs from general familiarity in that it is determined by the individual subject’s face-space rather than properties of the face itself.

Distinctiveness in Face Recognition

It has been shown many times that the distinctiveness of a face affects its recognition (see Hosie and Milne, 1995, for a detailed overview of this area). One robust observation is that distinctive faces are more likely to be correctly recognised than typical faces (Bartlett et al., 1984; Bruce, Burton, & Dench, 1994; Hancock et al., 1995; Light et al., 1979; Shepherd, Gibling, & Ellis, 1991; Vokey & Read, 1992). A second finding is that familiarity decisions can be made faster to distinctive faces than to typical faces (Johnston & Ellis, 1995; Valentine & Bruce, 1986; Valentine, 1991). A simple, almost trivial, explanation for this could be that distinctive faces are more likely to be responded to positively than typical faces. However, if this were the case, then one would expect to see more false positive errors for more distinctive faces as well as fewer miss errors for those faces. Light et al. found the reverse—distinctive faces elicited less false positive errors than typical faces. This research led to a depth of processing explanation which claimed that distinctive faces are processed to a deeper degree and so are remembered better or disregarded as distractors more easily.

Valentine (1991) offered a framework within which these distinctiveness effects may be explained and which can also account for the distinctiveness effects seen in face categorisation and the effects of inversion of faces. This framework was based on a multi-dimensional face-space in which a person encodes faces according to facial dimensions. These faces are stored as points or vectors in the space so that they can be retrieved in recognition. Within a person’s face-space, faces will tend to be distributed normally along each facial dimension, resulting in a space with a higher concentration of exemplars towards the centre. During recognition, the stimulus representation is compared with the stored exemplars and the one that is closest in the face-space will be the identity recalled. This means that distinctive faces will be easier to remember than typical faces because they can be encoded into a region of the person’s face-space with fewer exemplars. Recognition will be
easier for these faces because there will be fewer exemplars similar to the stimulus to compete for activation.

Valentine and Ferrara (1991) demonstrated how the face-space framework could be implemented using auto-associative and back-propagation neural networks to achieve a model for the effects described. This framework could also be modelled by the Interactive Activation and Competition with Learning (IAC) model suggested by Burton (1994). The face-space framework is also implied by the auto-associative memories implemented in the PCA-based models (e.g. Craw, 1995; Hancock et al., 1996; O’Toole, Deffenbacher, Valentin, & Abdi, 1994). In these models, the facial dimensions are extracted as eigenfaces. All these implementations offer a mechanism by which recognition is inhibited by the presence of stored faces which are similar to the stimulus but are not actually the target face. In this way, they can model the competition for activation between stored exemplars that are similar to the stimulus representation and so account for distinctive faces being easier to recognise.

The face-space framework has also been supported by empirical evidence. Bruce et al. (1994) examined how the distinctiveness of 175 faces affected both false positive errors and correct acceptance of targets (hits) in a recognition memory experiment. They found that distinctive faces received fewer miss errors than typical faces, and that distinctive distractor faces produced less false positive errors than typical distractors. Both these results are consistent with the findings of Light et al. (1979). However, Bruce et al. extended their analysis and found that although distinctiveness was found to correlate negatively with both types of errors, miss errors and false positives had a low correlation with each other. This result is particularly puzzling, since if a factor predicts two variables, then one would expect those variables to have some level of correlation that reflects their common predictor. The experiment presented below attempts to address this anomalous result.

The explanation considered is that miss errors and false positive errors are determined by different mechanisms but are both predicted by distinctiveness, either directly or indirectly. It has been postulated that distinctiveness predicts miss errors because the probability of producing a hit is determined by the number of surrounding stored exemplars in a person’s individual face-space. A new face that has been encoded into a region of the face-space with many similar known exemplars will not be encoded as well as one which is encoded away from close exemplars. Valentine’s (1991) face-space framework gives this kind of account for distinctiveness effects. On this account, the general distribution of exemplars in the subjects’ individual face-spaces are influential in determining miss errors. The general distribution of face-spaces will be similar between subjects, which means that the number of surrounding stored exemplars near a
face representation, and hence the distinctiveness, will be consistent between subjects.

A different mechanism is postulated for the production of false positive errors. A false positive error occurs when a subject considers the specific familiarity of the face to be high, even though the face has not been seen. One possible reason for a high specific familiarity is if the personal (or general) familiarity of the face is high and the context is confused. So, if the distractor face looks like a previously known face, then the personal familiarity will be higher and the probability of a false positive being produced is increased. This type of mechanism is alluded to by Vokey and Read (1992).

According to this account, false positive errors are determined by the proximity of a single particular stored exemplar. The closer the face representation is to a known face, then the more likely it is that a positive response will be produced. If the face is known, then the positive response is correct and so close proximity to the known exemplar is an advantage. If the face is unknown, then the close proximity to the stored exemplar will increase the probability of producing an erroneous positive response and hence a false positive error. This mechanism predicts that false positive errors are highly sensitive to the idiosyncratic differences between the position of exemplars in a face-space because they are determined by the proximity to only one stored face and not the general distribution of faces. Due to the normal distribution of faces, distinctiveness will negatively predict false positives because unfamiliar typical faces are more likely to be represented near a known exemplar than unfamiliar distinctive faces.

These two proposed mechanisms suggest that the rated distinctiveness of a face will predict both false positives and hits. However, the predictive power of this factor is based on general agreement between the distribution of exemplars in subjects’ face-spaces and is mediated via the mechanisms described. These mechanisms would also predict that if individual ratings of faces were used, instead of ratings averaged over many subjects, then a different pattern of predictors would occur. In this case, the idiosyncratic effect of personal familiarity could be uncovered. These mechanisms suggest that personal familiarity should be found to predict false positives, whereas distinctiveness would predict miss errors.

Inter-subject Consistency

Hancock et al. (1996) replicated the procedure of Bruce et al. (1994) and found similar results. In their analysis, they considered the degree to which subjects concurred in their distinctiveness ratings of the faces and
whether subjects tended to produce errors on the same faces or not. They found that there was good inter-subject agreement for the distinctiveness ratings: Subjects tended to agree on which faces were the distinctive ones. Inter-subject consistency for which faces elicited miss errors was lower than the agreement on the distinctiveness ratings, although this difference was not statistically analysed. This finding is interesting because it implies a greater level of idiosyncrasy between subjects for the hits rather than for the distinctiveness ratings. Such a combination of results would occur if subjects made the rating of distinctiveness by abstracting the difference between the central tendency of a face-space and each presented face. Since the general structure of the face-space would be similar for all subjects, the central tendency of each subject’s face-space will also be similar. This fact means that subjects would tend to agree on distinctiveness if they were extracting this measure from their knowledge about the general distribution of faces rather than from the exemplars they are familiar with themselves.

The mechanism for producing hits, mentioned above, is based on the pattern of exemplars within individuals’ face-spaces. If there are many known exemplars neighbouring the new face, then encoding of that face will be more difficult. This means that the idiosyncratic distribution of known exemplars within a subject’s face-space will affect how well the new face representation is stored. Due to the general distribution of the population of faces, there would be some agreement on which faces are encoded better, but this agreement would not be as high as when subjects use general distribution information, as it is suggested occurs for the distinctiveness judgements.

Using this interpretation of the results of Hancock et al. (1996), rated distinctiveness is a distribution-based measure that is relatively independent of idiosyncratic differences. Encoding a new exemplar and hence the production of hits is dependent on idiosyncratic differences in the distribution of exemplars in a subject’s face-space. Since each individual’s face-space is sampled from the general population of faces, these will approximate to the normal population and hence give rise to the correlation between hits and distinctiveness.

It is also possible to consider the inter-subject consistency of false positive errors. This is dealt with in detail below.

Personal Familiarity in Face Recognition

Bartlett et al. (1984) considered the mechanism through which context-free familiarity could predict face recognition and in particular the production of false positive errors. This mechanism was based on three premises. The first is that every face encountered will have a non-zero
level of familiarity. This is what is meant by personal familiarity as defined here, since it includes familiarity to faces that have not been seen previously. Also, typical unseen faces have a higher level of familiarity than distinctive unseen faces. The second premise is that exposure to a face will increase its familiarity. This premise accounts for the learning that takes place as a face becomes known and hence refers to specific familiarity, as does the third premise. The third premise is that, for equivalent presentation, the amount a face’s familiarity increases depends on the distinctiveness of that face. Distinctive faces have a faster increase in familiarity than typical faces. This point accounts for the fact that, although typical faces tend to start off with higher levels of familiarity when novel, distinctive faces have higher familiarity after the faces are learnt.

The effects of familiarity have been investigated in a variety of ways. Vokey and Read (1988) and Bartlett and Fulton (1991) investigated the role of subjective familiarity (a similar property to personal familiarity) on the production of false positives. Their designs were similar and used no target set but just one set of distractors. Subjects were required to judge if any of the faces reappeared in a series when in fact each face was presented only once (i.e. specific familiarity). Bartlett and Fulton’s design differed in having subjects who were either young or old. For their younger set of subjects, they found that distinctiveness predicted false positives whereas subjective familiarity did not. Vokey and Read found a correlation between distinctiveness ratings and subjective familiarity ratings, but Bartlett and Fulton did not find this to be significant. Bartlett and Fulton’s explanation of their results included the context-recollection hypothesis. The correlation between false positives and subjective familiarity seen in their older subjects is due to subjects’ deficiency in encoding context and so new faces are confused with both old test faces and previously known faces.

Vokey and Read (1992) analysed the composition of facial distinctiveness and found that it can be partitioned into two factors that differentially affect recognition. These factors are memorability and familiarity. In support of previous claims about distinctiveness, they found that distinctiveness correlated positively with hits and negatively with false positives. They employed multiple regression using the two rated properties of memorability and general familiarity to predict how subjects (from 10th and 12th grades) would perform in a recognition task with various faces as targets and distractors. They found that false positives were predicted by high general familiarity ratings and by low memorability ratings. The predictors of hits appear to be less clear-cut, with memorability being a positive and a negative predictor depending on whether the test used identical or different views of the target faces.
To interpret context-free familiarity in terms of the face-space, Bruce et al. (1994) suggested that it can be seen as the proximity of a novel face representation to other existing faces’ representations and so is related to personal familiarity. This theory would imply that more distinctive novel faces will be seen as less familiar due to the population distribution of face representation. A typical stimulus face is more likely to have a known representation close to the stimulus representation than a distinctive stimulus face. This would account for the correlation between averaged distinctiveness and familiarity observed by Vokey and Read (1988, 1992; see also Vokey and Read, 1995) and is consistent with the first premise described by Bartlett et al. (1984).

From the results of their experiments, Hancock et al. (1996) confirmed that there are two different factors affecting recognition errors. By a process of factor analysis they found that one factor, labelled memorability, positively correlated with distinctiveness ratings and probability of hits. This factor was also negatively correlated with the probability of the production of false positives. The second factor, labelled familiarity, correlated with false positives and with hits. The label of familiarity was used although the faces were not rated on their familiarity and so this label was based on the assumption that false positives were produced because of high familiarity. The experiment below examines this assumption as well as looking at other factors in the recognition of faces.

It has been pointed out (P.J.B. Hancock, personal communication) that the results of the factor analysis reported in Hancock et al. (1996) are not consistent with the mechanisms suggested here. If the second component is familiarity which takes account of the variation in the false positive performance, why do false positives correlate with the first principal component? There are two possible reasons for this. The first is that factor analysis uses averaged scores while personal familiarity is suggested to predict false positives only at the individual level. By averaging over many subjects, the subject-specific familiarity and the general distinctiveness of faces will become mixed up. The second reason for this correlation of the first principal component and false positives is that the orthogonal factors are not appropriate to extract memorability and familiarity. Vokey and Read (1995) demonstrate how these two factors should be considered to be oblique rather than orthogonal. Preliminary re-analysis of results reported by Hancock et al. suggest that an oblique factor analysis would remove the correlation between false positives and the first principal component.

There is a further result found by Hancock et al. (1996) which does not support Bruce and co-workers’ (1994) account of false positives. They found that there was a high degree of inter-subject consistency for false positive responses: The same faces tended to elicit erroneous “seen
before" responses more than other faces and this pattern was fairly consistent across subjects. As discussed above, a high degree of inter-subject consistency implies that there is a general property of those faces that elicit the response. In the case of the distinctiveness ratings, it was speculated that some knowledge about the relationship of a face's representation to the population of faces is involved in the mechanism employed by subjects. This high inter-subject consistency for false positives is contrary to the explanation of false positives as being a result of the stimulus face representation's similarity to one other known face. This explanation would predict low levels of subject agreement because false positives will be influenced by the idiosyncratic structure of each subject’s face-space. Attempting to reconcile the data from the experiments by Hancock et al. with the theory of production of false positives is an additional aim of the present study.

**EXPERIMENT**

As detailed in the previous sections, there are three main goals of this experiment. The first is to explain why Hancock et al. (1996) found a high level of inter-subject consistency for false positives. The second goal is to examine the relationship between the distinctiveness and personal familiarity of faces, and errors produced in a recognition paradigm. The third goal is to attempt to understand why hits and false positives have low correlations even though distinctiveness predicts both.

The experiment employed a face recognition paradigm similar to that used by Hancock et al. (1996). In the experiments reported by Hancock et al., the faces were split into two groups, one that was used as a training set and one that was used as a distractor set. These sets were interchanged between subjects. In their design, the distractor items were compared with the same target set for all subjects. This means that if a distractor face is similar to a target face, then this distractor face will tend to elicit false positive responses across different subjects. Here it is considered if such similarities between faces of different sets could account for the high level of agreement on false positives observed. To do this, the faces in this experiment were split into three groups so that each group of distractors could be used with two different sets of targets. It was predicted that if the target faces affect erroneous selection of the distractor items, then the inter-subject consistency would be lower between subjects who saw different sets of targets than between those who saw the same target faces. If this is correct, then the high level of subject consistency for false positives seen by Hancock et al. (1996) was potentially due to the same target faces being used for all subjects.
Distinctiveness and personal familiarity of the faces were rated in order to examine their predictive power for recognition errors. Since the particular idiosyncrasies of subjects are of interest, it was necessary to have the same subjects both to rate the faces and take part in the recognition experiment. The rating task also acted as a learning session for the subjects. To obtain ratings for the distractor items, it was necessary to have these rated after the test phase was completed. It was hypothesised that distinctiveness of faces would positively predict hits and personal familiarity would positively predict false positives.

**METHOD**

**Subjects**

Thirty-six undergraduate students from the University of Wales received course credit for their participation in the experiment.

**Materials**

Sixty grey-scale photographs of men were used in the experiment. The images were all 3/4 views of the face with a neutral expression. None of the faces had facial hair, spectacles or other distinguishing features.

**Procedure**

The set of 60 faces was split up randomly into three sets of 20 (sets A, B and C). Each subject was presented with one set and was required to rate each face on its distinctiveness and its personal familiarity on two 10-point scales. A distinctive face was described as one “which stands out from a crowd”; at the other end of this scale were faces “which would be hard to recognise in a crowd”. To obtain a rating of personal familiarity, the subjects were asked whether they thought that the face looked similar to (or reminded them of) someone they knew previous to beginning the experiment. The end-point of the scales were described as “a face almost identical to someone you know” and “like no face you have ever seen before”. Half the subjects rated the faces for distinctiveness then personal familiarity and the other half did the rating in the reverse order. The presentation of each face was for 5 sec, but subjects could take as much time as they required to make their decision. The faces in each rating procedure were presented in a random order. Ratings on the two scales took place in two separate blocks, so that each subject saw each face for up to 10 sec.

After the rating procedure was completed, the subjects performed an unrelated text memory task, which lasted approximately 5 min. Immedi-
ately following this, subjects were presented with an unexpected face recognition task. Two intermixed sets of faces were presented randomly to the subjects. One of these was the set of faces they had rated (targets) and the other was one of the sets they had not seen before (distractors). The target and distractor faces were chosen so that all six possible combinations over the three sets were used six times over the 36 subjects. Subjects were required to decide whether each face had occurred in the rating part of the experiment or not. Each picture appeared for 5 sec but subjects could take as long as they required to make their decisions.

Finally, subjects rated the distractor items for personal familiarity and distinctiveness on the same 10-point scales used for the target faces. To obtain personal familiarity ratings, it was stressed to subjects that it was similarity to faces seen prior to starting the experiment that was required.

RESULTS

The results were processed in several different ways. The collected data were split up into rating data and recognition data. Rating data were the subjects’ ratings of distinctiveness and personal familiarity. Recognition data were whether subjects responded to targets and distractors correctly in the test phase. In general, the subjects performed well at the recognition stage. For the target faces, the probability of a correct response was 0.859 (SD = 0.110; n = 36); for the distractor faces, the probability of correct rejection was 0.857 (SD = 0.090; n = 36). Table 1 shows how the ratings and the recognition performance break down over the three sets of faces.

Correlations Between Ratings Scores

The mean distinctiveness and personal familiarity values of each of the 60 faces were calculated from the subjects’ ratings. These two factors were

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>The Means (± SD) of the Ratings and Recognition Performance on Faces Broken Down According to the Three Training Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distinctiveness</td>
</tr>
<tr>
<td>Set A</td>
<td>5.196 ± 1.088</td>
</tr>
<tr>
<td>Set B</td>
<td>5.479 ± 1.271</td>
</tr>
<tr>
<td>Set C</td>
<td>5.252 ± 1.105</td>
</tr>
</tbody>
</table>

n = 20 in each set.
found to have a significant negative correlation \[ r(58) = 0.393, P < 0.005 \]. If individuals’ ratings were used instead of mean ratings, then the correlation was not statistically significant \[ r(1438) = 0.032, P > 0.05 \].

Correlations Between Recognition Scores

The probability of each face being classified correctly either as a target or a distractor was calculated by finding the mean over all subjects in a particular condition. These two factors were found to have a non-significant negative correlation \[ r(59) = 0.083, P > 0.05 \].

Inter-subject Consistency

Following the procedure adopted by Hancock et al. (1996), inter-subject consistency was assessed to examine whether the same faces were rated as distinctive or familiar to a similar extent by all subjects and whether subjects tended to make errors on the same faces. For both the distinctiveness scores and the personal familiarity scores, each subject’s ratings were correlated against the ratings of one of the other subjects who had rated the same set of faces. Likewise, the pattern of responses to faces at the recognition stage was compared between subjects who were tested on the same targets or distractors. The ratings from each subject were used only once to ensure independence of the variable produced.\(^1\) In this way, the overall consistency of subjects’ ratings could be estimated by the means of the correlations between subjects. If subjects were consistent on which faces were distinctive or familiar, then this should appear as high correlations. Even if the level of consistency between subjects is low, by comparing between a large number of subjects it is possible to find whether the small consistencies are robust; moreover, it is possible to compare the size of these small correlations over different ratings. If the ratings are more related to the idiosyncrasies of each subject, then this will produce low correlations. If the ratings are affected by general properties of the faces, available to all subjects, then the correlations between subjects will be slightly higher.

\(^1\)A similar set of analyses was conducted by comparing each subject’s ratings with the ratings of the 23 other subjects who judged the same faces. Although this violates the assumption of independence, it was conducted to investigate whether the results of the analysis were an artefact of the sampled comparisons. These analyses found the same pattern of significance as the unbiased sampled analyses and so imply that the results described are not an artefact of sampling. The same procedure was conducted for the rating correlations, which also yielded the same results as the sampled analysis.
Distinctiveness and Personal Familiarity. The mean of the correlations between subjects’ distinctiveness ratings over all three sets of faces was 0.262 (SD = 0.264). The null hypothesis that the mean correlation between subjects is zero was tested using a one-sample t-test on the set of correlations. The correlation values were converted to $z'$ scores\(^2\) to deal with the skewed distribution of $r$ values whose mean is not zero (Cohen & Cohen, 1983). Analysis on the $z'$ scores gave a significant difference from zero [$t(32) = 5.362, P < 0.001$], implying that the average correlation between subjects is above zero even though the size of any individual correlation is below usual significance levels.

For the personal familiarity ratings, the mean of the inter-subject correlations was 0.055 (SD = 0.216). A t-test on the $z'$ scores found this to be non-significant [$t(32) = 1.403, P > 0.05$], so the mean correlation between subjects is not significantly above chance. Furthermore, a two-sample t-test on $z'$ scores found that the consistency of the distinctiveness ratings is significantly greater than the consistency of the personal familiarity ratings [$t(64) = 3.632, P < 0.001$]. It would appear that subjects agree more on which faces are distinctive than on which faces are familiar.

Hits and False Positives. The inter-subject consistency of hits was calculated by correlating the responses of each subject with those of another subject who had rated the same faces prior to testing. Subjects who made no “miss” errors were excluded due to the lack of variance in their data. Data from seven subjects were removed for this reason, which still left 14 correlations that could be used. Over the three sets of faces, the mean inter-subject correlation was found to be 0.174 (SD = 0.286), which was found to be significantly above zero using a t-test on $z'$ scores [$t(13) = 2.298, P < 0.05$].

For the false positives, each inter-subject correlation was classed according to whether the target set was the same for both subjects or different for both subjects. That is, if two subjects both rated set A and were tested on sets A and C, then this correlation is in the “same” group. Consider now one subject who rated set A and was tested on sets A and C, and another subject who rated set B and was tested on sets B and C. The false positives both occur for the faces in set C and so these two subjects’ false positive scores can be correlated. In this case, the correlation will be classed as “different”, since the subjects learnt different target faces. Again data from subjects that made no false positive errors were

\(^2\)The transformation $z' = \frac{\ln(1 + r) - \ln(1 - r)}{2}$ produces a sampling distribution that closely approximates the normal distribution (Cohen & Cohen, 1983).
excluded due to lack of variance. Two subjects were removed from the analysis for this reason.

For the “same” correlations, a mean of 0.217 (SD = 0.265) was found, which is significantly above chance level \( t(15) = 3.231, P < 0.001 \) using a \( t \)-test on \( z' \) scores. This meant that these subjects tended to get the same pattern of faces wrong from the distractor set if they had been trained on the same target set.

The “different” correlations produced a mean of 0.029 (SD = 0.272), which was not found to be significantly different to chance when a \( t \)-test was conducted on the \( z' \) scores \( t(15) = 0.536, P > 0.05 \). Furthermore, a two-sample \( t \)-test on \( z' \) scores showed that these inter-subject correlations were significantly lower than the correlations between subjects who saw the same target faces \( t(30) = 1.851, P < 0.05 \). Thus, it would appear that the inter-subject consistency is larger when the same target faces are used.

Distinctiveness and Personal Familiarity as Predictors

The predictive powers of the rated properties of faces on the errors they elicit were analysed using multiple regression in two ways. First, the probabilities of producing errors were regressed against the mean rating of distinctiveness and familiarity for each face using the technique employed by Vokey and Read (1992). Second, the individual ratings were used to predict individuals’ errors on particular faces. This analysis was conducted to investigate the idiosyncratic effects of subjects’ face-spaces.

**Averaged Ratings.** Two multiple regressions were conducted: one with the probability of hits as the dependent variable, and one with the probability of false positives as the dependent variable. For both regressions, the independent variables were the standardised personal familiarity and distinctiveness of each face averaged across all subjects who rated that face. An interaction term was also included in the analysis. The standardisation, or conversion to \( z \) scores, of the independent variables is necessary for an unbiased investigation of the interaction (see Aiken & West, 1991).

Distinctiveness was a significant predictor of probability of producing a hit \( r = 0.353, t(56) = 3.638, P < 0.001 \) Personal familiarity was not a significant predictor of hits \( r = 0.036, t(56) = 1.653, P > 0.05 \) nor was the interaction \( r = 0.029, t(56) = 1.498, P > 0.05 \). Removal of the non-significant terms did not affect the significance of the other terms. The regression equation with significant terms only is:

\[
p(\text{Hit} \mid \text{Target}) = 0.061 \times \text{Mean(Distinctiveness)} + 0.876
\]
where \( p(\text{Hit} \mid \text{Target}) \) is the probability of a face producing a hit when a target item is presented, and Mean(Distinctiveness) is the standardised mean distinctiveness of a face.

For the false positive regression, none of the factors—distinctiveness, personal familiarity or the interaction—was significant [distinctiveness: \( r = 0.131, t(56) = 1.012, P > 0.05 \); personal familiarity: \( r = 0.024, t(56) = 0.234, P > 0.05 \); interaction: \( r = 0.023, t(56) = 0.215, P > 0.05 \)]. Again the removal of the interaction term did not affect the significance of the other terms. A hierarchical analysis also found all the terms to be non-significant.

**Signal Detection.** An alternative way to interpret the relationship between rated properties of faces and recognition performance is to employ signal detection theory. It is possible to transform measures of hits and false positives into sensitivity (\( d' \)) and criterion. Two multiple-regression analyses were used to discover whether distinctiveness and personal familiarity predict the performance measures of \( d' \) and criterion. Twenty-seven of the 60 faces produced either no miss errors or no false positive errors. The \( d' \) and criterion for these faces could not be calculated. The remaining 33 faces were used in the analysis. For the \( d' \) regression, none of the factors was significant [distinctiveness: \( r = 0.170, t(29) = 1.367, P > 0.05 \); personal familiarity: \( r = 0.194, t(29) = 1.512, P > 0.05 \); interaction: \( r = 0.036, t(29) = 0.632, P > 0.05 \)]. Also, for the criterion regression, none of the factors was significant [distinctiveness: \( r = 0.049, t(29) = 0.015, P > 0.05 \); personal familiarity: \( r = 0.057, t(29) = 0.597, P > 0.05 \); interaction: \( r = 0.237, t(29) = 1.416, P > 0.05 \)].

The reason for these non-significant results may stem from the type of task employed. The calculation of \( d' \) and criterion has the implicit prerequisite that all trials are equivalent in terms of the relationship between signal and noise. It has already been noted that subjects have idiosyncrasies determined by the structure of their own face-spaces. This would imply that \( d' \) and criterion would vary between subjects as well as between faces, and so trials between subjects are not necessarily equivalent in terms of signal detection. This additional degree of variation may account for the non-significant results found. To avoid this problem, individual performance on particular faces was considered. This trial-by-trial consideration of the results means that it is not possible to employ the principles of signal detection.

**Individual Ratings.** Since there was a low degree of subject consistency for the personal familiarity scores, it is possible that individual differences will affect the overall predictive power of personal familiarity scores. To
incorporate this in the analysis, individuals’ ratings of faces were considered as predictors of individuals’ recognition of those faces. This analysis uses the dichotomous scores as the dependent variable in the regression. Although this is a violation of the formal model, Overall (1980; see also Tabachnick & Fidell, 1989) demonstrated how dichotomous dependent variables could be usefully employed in a least square multiple regression. Again the independent variables in the regression were standardised distinctiveness, standardised personal familiarity and the interaction between these. In this analysis, individual scores were used rather than averaged values.

For the hits, only distinctiveness was a significant factor \[r = 0.102, t(676) = 2.558, P < 0.05\] Personal familiarity did not reach significance \[r = 0.071, t(676) = 1.742, P > 0.05\] nor did the interaction between personal familiarity and distinctiveness \[r = 0.006, t(676) = 0.278, P > 0.05\] Removing the non-significant terms did not affect the significance of the main predictors. The regression equation derived from individual scores is of the form:

\[
\text{Hit} = 0.034 \times \text{Distinctiveness} + 0.859
\]

In the regression analysis of the false positives, using individual scores, only personal familiarity was a significant factor \[r = -0.085, t(616) = 1.991, P < 0.05\] The distinctiveness factor and the interaction were not significant \[r = -0.022, t(616) = 0.526, P > 0.05; r = -0.038, t(616) = 0.657, P > 0.05\] respectively. Removing the non-significant terms did not affect the significance of the main predictors. The regression equation using individual scores as predictors is:

\[
\text{FP} = 0.030 \times \text{Familiarity} + 0.143
\]

**Summary**

There is higher inter-subject consistency on ratings of distinctiveness than for ratings of personal familiarity. Also, it was found that these two factors were negatively correlated with each other when averaged across subjects. The inter-subject consistency for which faces received hits was above chance. The inter-subject consistency for false positives was not significantly above chance when subjects who had seen different targets were compared. If the subjects had seen the same target faces, then their false positives tended to occur on the same faces.

Multiple regression was used to consider the way the rated values of faces predict the errors produced. Using grouped means of rating and recognition data, we found that the probability of hits increased as faces became more distinctive. No predictors were found to have a significant effect on the number of false positives produced. Using individual ratings...
to predict individual errors produced slightly different results. In this case, hits were still increased significantly for the more distinctive faces and personal familiarity was not a significant predictor of hits. Individuals' personal familiarity ratings were found to be a significant predictor of whether that individual would produce a false positive on a given item. Faces with higher personal familiarity ratings were more likely to yield a false positive. Using individual scores, the predictive power of distinctiveness on false positives was non-significant.

**DISCUSSION**

The results are considered in three ways in line with the three main goals described above. First, the effect of changing the target face while keeping the same distractors is considered. This examines the role that the targets play on the false positive errors. Second, the relationship between the rated properties of faces (distinctiveness and personal familiarity) and the types of errors produced are examined. Third, it is considered how these results help to account for why it has been found that false positives and hits do not correlate, even though both have been found to be predicted by distinctiveness.

**Target Face Effects**

Hancock et al. (1996) report that the consistency between subjects for the errors they produce is greater for false positives than for miss errors. In attempting to explain this result, they suggest that the false positives demonstrate a general property of the faces, whereas miss errors reflect idiosyncratic differences between subjects. The results reported here mirror these results. For miss errors, the mean correlation between subjects was 0.175. When subjects had seen the same target set (as they always did in the experiments reported by Hancock et al.), the mean correlation between subjects for false positives was 0.217, which is greater than that for miss errors.

The interesting difference between our results and those reported by Hancock et al. (1996) is what happens if subjects who rated different targets but were tested on the same distractors are compared. In this case, the mean correlation drops to 0.029, which is lower than that found for hits. This result means that the false positives produced by subjects were influenced by the particular targets seen previously and so the set of target faces has an effect on the selection of faces when assessing the distractor items.

Hancock et al. (1996) were correct when they accounted for the higher consistency for false positives than for miss errors as a general property
of the faces. The present experiment provides evidence about the nature of this general property. It appears to be the relationship between the distractors and the particular target sets of faces that produce the higher levels of consistency. The suggestion that “false positives arise as a consequence of a face’s similarity to the general population” (Hancock et al., 1996, p. 39) can be rejected as an explanation, since all subjects have face-spaces sampled from the same population, but the consistency drops if the target set is not the same.

It had previously been considered that the targets had little effect on the false positives and the determining factor was the general familiarity of the faces; after all, the target face set contains very few exemplars relative to the hundreds of known faces a subject brings to the experiment. This experiment found that personal familiarity is a determining factor of false positives, but the size of the difference produced by changing just the 20 target faces is worth consideration separately to the effects of distinctiveness and personal familiarity of the faces.

In some way, false recognition of distractors is influenced by the small training set, even though the subject may know hundreds or thousands of other faces. Two mechanisms by which the small set of target faces could disproportionately affect incorrect selection of distractor faces are considered. These are the effect of recency of presentation, and the effect of the learning context on recognition.

During the test phase, the target set contained those faces most recently seen by the subject. This means that these faces are primed and in a higher state of activation, to use an interactive activation model as a metaphor (e.g. Burton’s, 1994, IACL model). This heightened level of activation may increase their influence in the familiarity decision. All other known faces are in a relatively low activation state, since they have not been seen recently. These low-activation exemplars will produce the personal familiarity effects as described below but the more recently seen faces will have an effect due to their heightened activation.

The second explanation is that the target faces were presented in a unique situation. This set is linked by its context and is distinguishable from other face traces due to this context. In this case, all faces in the experiment fall into the “seen-in-this-experiment” context. Since the distractor faces were presented in the same context, these target faces would have a greater effect. This can be interpreted as the context being an extra dimension in the face-space and thus the distractor items will be more similar to the targets (i.e. same context) than the other known faces [such an explanation could be modelled using Valentine and Ferrara’s (1991) framework with the inclusion of extra “context dimensions” into the face patterns] The personal familiarity effect still occurs because it is free of context, but since the target faces have this disproportionate
effect, an element of context should be incorporated into theories of false positive error production.

Regardless of whether these two accounts are correct, it is still the case that the target set influences the production of false positives. This does not rule out the fact that personal familiarity of the distractor items may also affect false positives, but it does mean that it is necessary to consider the target faces when interpreting what does and does not affect false positives. The next section deals with how the rated properties of the faces predict how subjects perform in the recognition task.

**Distinctiveness and Personal Familiarity as Predictors of Recognition**

The inter-subject consistency of the distinctiveness ratings is higher than the consistency of the personal familiarity ratings. This is concordant with the idea that distinctiveness ratings are based on knowledge about the distribution of faces, whereas personal familiarity is due to the similarity to particular known exemplars. The general distribution of faces will be approximately the same for most subjects, so there will be high agreement for measures based on this, whereas similarity to known exemplars, and hence personal familiarity, will be more sensitive to idiosyncratic differences between subjects. Within a normally distributed space, these two factors would be related when averaged across subjects, since this removes the idiosyncrasies of subjects. A face that is close to the central tendency is more likely to have another exemplar face near it than a face a long way from the central tendency. This fact is represented in the data by the negative correlation between distinctiveness and personal familiarity when averaged across subjects and a non-significant positive correlation when the individual ratings are compared.

Using multiple regressions, it was found that distinctiveness ratings predicted the number of hits produced for a face. The more distinctive a face, the more likely it was to be identified correctly as being seen before. This effect was significant regardless of whether or not the regression was conducted on the scores and ratings that had been averaged across subjects.

This result is consistent with the idea that, for a face to be remembered, it is better to have a relatively sparse part of the face-space on which to encode it. If it has many similar neighbours, then retrieval (and possibly encoding) will be more difficult.

The multiple regression using mean ratings as predictors of false positives gave no significant effects. It is possible that the effect of personal familiarity is being masked because using mean values hides the
idiosyncratic differences between subjects that may be of more importance as a predictor. This possibility is supported by the regression using individual ratings to predict individual responses. In this case, the personal familiarity of the face predicted the number of false positives produced. The more similar the face was to someone the subject knew (i.e. personal familiarity), the more likely it was that the face would produce a false positive response.

It appears that the false positives are affected by the idiosyncrasies of individuals’ face-spaces. This may be because they are affected by the presence of just the closest neighbour. In Burton’s (1994) competitive recognition system, mentioned above, a novel face is more likely to produce a feeling of familiarity if it looks like just one known face. In this way, there is little competition from other possible candidates to reduce the face activation. This means that any benefit offered to a typical face, in that it is more likely to be close to an encoded face, will be removed by the other close neighbours, which will also be partially activated. In this way, the face most likely to produce a false positive is one which has a representation very close to one encoded representation while being a long way from any others. It is possible that this relativity of proximity is going to be unaffected by the distinctiveness of a face. A distinctive face is as likely to have a high relative proximity than a typical face. This can also account for why a correlation between faces that produce high levels of hits and those that produce high levels of false positives is always found to be weak. However, it is just as likely that the distance of the closest exemplar outweighs the other relative distances in the calculation of relative proximity. In this case, false positives would occur more often for typical faces than distinctive faces.

**Hits and False Positives**

To return to the question asked above, why is it that although distinctiveness predicts both hits and false positives, no correlation has been found between these responses? First, it is worth noting that distinctiveness did not predict false positives in this experiment even though this relationship has been found before. The pattern of results reported here shows that personal familiarity and distinctiveness negatively correlate when they are averaged over many subjects. Also, distinctiveness of a face predicts the hits it receives, whereas personal familiarity predicts false positives but only when individual scores are used. It has also been demonstrated that if the effect of having the same target set is removed, then the idiosyncrasies in false positives are larger than for hits. These results can be put together to account for why no correlation has been found between hits and false positives.
Hits do not correlate with false positives, although distinctiveness can be used as a predictor of false positives, because the relationships are not a direct cause-and-effect and moreover there are two different mechanisms involved. Distinctiveness is a predictor of personal familiarity if it is summed over many subjects: Distinctive face representations will tend to occur in regions of subjects’ face-spaces that contain many other exemplars, thus producing high personal familiarity ratings. Personal familiarity is a predictor of false positives but only when individual scores are used. A face that is familiar to someone is likely to elicit a false positive in that person, but that is no guarantee that another subject will also give a false positive on that face. These points mean that although distinctiveness can predict both hits and false positives (as seen in Hancock et al., 1996), the mechanisms involved are very different. Distinctiveness predicts which faces elicit miss errors because faces close to the centre of the face-space are encoded in a dense region of the face-space making retrieval problematic. Distinctiveness predicts which faces elicit false positives because a novel face is more likely to be similar to a known exemplar (or even a target item) if the novel face occurs towards the centre of the face-space. These different mechanisms mean that the correlation between the two types of errors is going to remain small even though distinctiveness can be seen to be a predictor of both.

CONCLUSION

The present experiment differs from other face recognition experiments in two important ways. The first is that the faces were rated by the same subjects as took part in the recognition task; moreover, each subject provided ratings for targets and distractor items. This is important, since it allowed the analysis of performance as predicted by both general rated measures and also the individual measures given by each subject for each face. Manipulating whether mean or individual scores were used made little difference to the predictive power of distinctiveness, but it was vital for the prediction of false positives by personal familiarity. This result leads to the conclusion that distinctiveness acts as a result of the general property of the face, whereas personal familiarity and hence false positives are more susceptible to subjects’ idiosyncrasies.

The second way this experiment differs from traditional face recognition experiments is the way the faces were split into three groups rather than two. This manipulation allowed the consistency on distractors to be compared between subjects who saw the same target set and between subjects who saw different target sets. It was found that the high inter-subject consistency for false positives reported by Hancock et al. (1996)
was a result of the same items always being used as targets for each distractor set. This implied that the target faces have a disproportionate effect on the false positives compared with all the other faces a subject knows.

The findings from this experiment suggest that personal familiarity is a useful predictor of which faces elicit false positives, but this is most appropriate when considered on an individual by individual basis. Also, the results show that the target items used may greatly influence the false positives produced and so any experiment used to investigate these errors should take this into account.

Manuscript received October 1996
Revised manuscript received March 1997

REFERENCES


