

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/3834>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



**Eyewitness Identification:
Improving Police Lineups for Suspects with Distinctive
Features**

By

Theodora Zarkadi

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

November 2009

Dedication

To the memory of my grandmother

Table of Contents

Table of Contents	i
List of Tables.....	iv
List of Figures	v
Acknowledgments.....	viii
Declaration	ix
Note on inclusion of published work	x
Abstract	xi
Introduction	1
Study 1	32
The Role of Distinctive Features in Face Recognition	32
Abstract.....	32
Introduction.....	33
Experiment 1	40
Method	40
Participants.....	40
Stimuli.....	40
Procedure	41
Results.....	43
Manipulation Check.....	43
Analysis of FAs to New Faces	43
Analysis of Hits to Old Faces	43
Signal Detection Analysis.....	45
Experiment 2.....	48
Method	48
Participants.....	48
Stimuli.....	48
Design and Procedure	48
Results.....	48
Analysis of Hits to Old Faces	49
Signal Detection Analysis.....	50
Experiment 3.....	52
Method	52
Participants.....	52
Stimuli.....	52

Design and Procedure	52
Results.....	52
Analysis of FAs to New Faces	52
Analysis of Hits to Old Faces	53
Signal Detection Analysis.....	54
Modelling.....	57
A Comparison with Deffenbacher et al.....	67
Locus of effect	68
Study 2	70
Creating Fair Lineups for Suspects with Distinctive Features.....	70
Abstract	70
Introduction.....	71
Stimuli.....	74
Experiment 1	75
Method	75
Participants.....	75
Procedure	76
Results.....	77
Experiment 2	78
Method	79
Participants.....	79
Procedure	79
Results.....	79
Discussion.....	81
Study 3	84
Lineup Construction for Suspects with Distinctive Features: To Replicate, Remove, or Pixelate?.....	84
Abstract.....	84
Introduction.....	85
Experiment 1	89
Method	89
Participants.....	89
Stimuli.....	89
Procedure	90
Results.....	94
TP lineups	95

TA lineups.....	95
Experiment 2.....	96
Method.....	97
Participants.....	97
Design.....	97
Materials.....	97
Study phase.....	100
Test phase.....	101
Results.....	101
Identification performance.....	101
Confidence-Accuracy Relationship.....	103
Study 4.....	114
Testing Police Lineups for Suspects with Distinctive Features Using a Videotaped Simulated Crime.....	114
Abstract.....	114
Introduction.....	115
Method.....	118
Participants.....	118
Design.....	118
The Video.....	118
The Lineups.....	119
Procedure.....	122
Results.....	123
Identification Performance.....	123
Confidence-Accuracy Relationship.....	125
Pre-lineup Confidence.....	125
Post-lineup Confidence.....	125
Change in Confidence.....	127
General Discussion.....	135
Summary and Discussion of the Experimental Findings.....	135
Theoretical and Practical Implications.....	141
Limitations.....	142
Future Directions.....	145
Conclusion.....	152
References.....	153

List of Tables

Study 1

Table 1. The Four Possible Responses to Each Face Stimulus.....	43
Table 2. Summed Similarities for the HS Model.....	59
Table 3. Best-Fitting Values of the HS Model's Parameters for 1-, 5-, and 10-Second Exposure Duration.	60

List of Figures

Introduction

Figure 1. Example of a biased lineup.....	10
Figure 2. Example of an unbiased lineup.....	11
Figure 3. A face (a) before and (b) after the digital alteration to include baldness, a blemish, a moustache, a beard (c) and glasses	15
Figure 4. Concealment is achieved by either (a) pixelating the area of the distinctive feature, or (b) covering the area of the distinctive feature with a black rectangle	16

Study 1

Figure 1. Examples of normal faces (top row) and the same faces after the digital addition of a mole, facial hair, a bruise and a scar (bottom row, from left to right) as presented in Experiments 1 and 2.	41
Figure 2. Mean proportions of Yes responses to (a) new faces and to (b) old faces in Experiment 1	45
Figure 3. (a) d' and (b) C as a function of study and test format in Experiment 1	46
Figure 4. Mean proportions of Yes responses to (a) new faces and to (b) old faces in Experiment 2	49
Figure 5. (a) d' and (b) C as a function of study and test format in Experiment 2 ...	51
Figure 6. Mean proportions of Yes responses to (a) new faces and to (b) old faces in Experiment 3	54
Figure 7. (a) d' and (b) C as a function of study and test format in Experiment 3	56
Figure 8. The HS model's predictions for (a) 1-second, (b) 2-second, and (c) 3-second exposure duration.....	64

Study 2

Figure 1. Examples of faces used in Experiments 1 and 2 before (top) and after (bottom) the digital addition of a distinctive feature (from left to right: a bruise, a mole, a piercing, a moustache, a scar, and a tattoo).....	75
Figure 2. Examples of (a) a replication lineup and (b) a concealment lineup presented in Experiments 1 and 2.	77
Figure 3. Mean proportion of correct responses and errors for replication and concealment lineups in Experiment 1	78
Figure 4. Mean proportion of correct responses and errors for replication and concealment lineups in Experiment 2: (a) target-present lineups and (b) target-absent lineups	80

Study 3

Figure 1. Example face used in Experiment 1 before (from left to right), after the digital addition of a tattoo, and after the pixelation of the area of the tattoo.	90
Figure 2. Examples of a (a) replication-technique lineup, (b) a removal-technique lineup, and (c) a pixelation-technique lineup presented in Experiment 1.....	93
Figure 3. Mean proportions of correct responses and errors under replication, removal, and pixelation for (a) TP lineups (a) and (b) TA lineups.....	94
Figure 4. Photo of the confederate in Experiment 2.	98
Figure 5. The two lineup techniques tested in Experiment 2: (a) replication, and (b) pixelation.....	100
Figure 6. Proportion of participants who made target identifications, foil identifications, and no identifications as a function of lineup technique (replication vs. pixelation) in (a) TP lineups and (b) TA lineups.....	102

Figure 7. Participants' percent distribution of self-reported pre-lineup confidence.	104
Figure 8. Participants' percent distribution of self-reported post-lineup confidence for replication and pixelation in (a) TP lineups and (b) TA lineups	106
Figure 9. Participant's percent distribution of change in self-reported confidence before versus after viewing (a) a TP lineup and (b) a TA lineup.....	108

Study 4

Figure 1. Photo of the confederate in Experiment 1.	119
Figure 2. The three lineup techniques tested in Experiment 1: (a) Replication, (b) Removal, and (c) Pixelation.....	122
Figure 3. Proportions of correct responses and errors in replication, removal, and pixelation lineups for (a) TP and (b) TA lineups	124
Figure 4. Participants' percent distribution of self-reported post-lineup confidence for replication, removal, and pixelation in (a) TP and (b) TA lineups.....	127
Figure 5. Participant's percent distribution of change in self-reported confidence for replication, removal, and pixelation before versus after viewing (a) a TP lineup and (b) a TA lineup	129

Acknowledgments

I would like to acknowledge my supervisors, Neil Stewart and Kim Wade, for their help and support during the completion of my PhD. I would also like to acknowledge Avraham Levi, Will Matthews, Zach Estes, and Rob Nash for criticisms and suggestions on chapters of this thesis as well as all of the participants who volunteered to take part in one of my studies. Finally, I wish to thank my husband for his encouragement every step of the way.

Declaration

I hereby confirm that I completed this thesis independently, that I have not heretofore presented this thesis to another department or university, and that I have listed all references used, and have given credit to all additional sources of assistance.

Note on inclusion of published work

Study 2 of this thesis has previously been published during the period of my PhD registration, and the copyright of this paper resides with the publishers (the reproduction of the paper as chapter in this thesis is permitted the terms of the copyright agreement). This paper is the following:

Zarkadi, T., Wade, K. A., & Stewart, N. (in press). Creating fair lineups for suspects with distinctive features. *Psychological Science*.

Study 3 is currently under review.

Abstract

Eyewitnesses' descriptions of suspects often refer to distinctive facial features, such as tattoos or scars, and the police have to decide how best to create fair lineups in these circumstances. This issue, despite its importance, has attracted insufficient attention in the eyewitness identification literature. Informed by the Police and Criminal Evidence Act code of practice and current police practice, I conducted an empirical evaluation of the different lineup techniques that investigators currently use for suspects with distinctive features.

To ensure that a suspect does not stand out because of his distinctive feature, and also to extract more information from the eyewitness, the police either replicate the distinctive feature across all foils in the lineup or conceal the distinctive feature on the face of the suspect. These techniques were tested either in a crossover recognition-memory paradigm (Study 1), or in a lineup-identification paradigm (Studies 2, 3, and 4), either in computer-based laboratory experiments or real-world field experiments using both target-present and target-absent lineups. The results showed that replication is a better technique than concealment. Compared to concealment, replication increases target identifications in target present lineups—in some cases by decreasing foil identifications in target-absent lineups. The hybrid-similarity (HS) model of face recognition was used to assess whether it could be applied in this domain. Across seven experiments (Studies 1, 2, and 3) and three paradigms, the HS model was able to model the qualitative pattern of results.

The purpose of this experimental work was to demonstrate the importance of constructing fair lineups for people with distinctive features and to provide results that will have practical implications for legal contexts and will improve our understanding of face recognition and recognition memory in general.

Introduction

Establishing the identity of the perpetrator is a prerequisite of every criminal prosecution in the courts of law. Unless the defendant confesses that he committed the criminal act, the prosecution will have to provide evidence that the defendant is the actual perpetrator and prove it at trial.

Dennis (2007) distinguishes between four types of evidence in courts of law. One type is the expert scientific opinion confirming that trace physical evidence (e.g., fingerprints, blood, hairs, samples of handwriting etc.) found in the scene of the crime matches material obtained from the defendant. A second type of evidence is circumstantial evidence (e.g., when the defendant cannot provide a good reason why he was at the scene of the crime, or when the investigation reveals that he had a motive to commit the criminal act that he is accused of). A third type of evidence is real evidence (e.g., evidence from CCTV or voice recorders that match the appearance of the defendant). Evidence of any of the above kinds does not by itself prove guilt of the defendant, but it is certainly an important factor to be taken into account. The fourth type of evidence is eyewitness testimony; this type of evidence, although it may exist only in the mind of the eyewitness, is direct evidence of guilt. Eyewitness testimony is used either as a supportive, additional form of evidence, or in the absence of other types of evidence.

Turtle, Lindsay, and Wells (2003) draw some direct comparisons between physical and memory trace evidence, adopting the so called “trace-evidence” approach to studying eyewitness memory. The term was initially introduced by Wells (1995) to describe eyewitness memory as something that the perpetrator left behind after he departed the scene of the crime, just like physical trace evidence. Memory traces can be delicate, hence –if not handled with care– easily damaged, can

decay over time and they can easily be cross-contaminated (e.g., when witnesses interact and share information just like blood from one area of the crime scene can be mixed with blood from another area during collection). Furthermore, the way that the eyewitness' memory is tested (e.g., the procedure of a lineup) can influence the reliability of the results just like the method that one uses to test physical evidence (e.g., a DNA profile) can have an effect on the reliability of the results. Finally, just as physical evidence can be fabricated (e.g., fake DNA, Frumkin, Wasserstrom, Davidson, & Grafit, in press), memory trace evidence can also be engineered when an eyewitness commits perjury.

Nevertheless, eyewitnesses exert a huge influence on the direction of criminal investigations and on the outcomes of trials. Eyewitnesses may be able to describe (parts of) the event that they witnessed and/or identify the offender who committed the relevant criminal act. The latter type of eyewitness testimony, that entails identification evidence, is the focus of this thesis. In most cases, after a positive identification is obtained from the eyewitness, the prosecution will most likely take the case to court (Tredoux, Meissner, Malpass, & Zimmerman, 2004). Once the case has proceeded to court, judges and juries must decide, sometimes solely on the basis of the identification evidence, whether the defendant is guilty. In fact, in many cases identification has been used as the only direct evidence of the guilt of the defendant (Wells & Olson, 2003). It seems that most of us, including jurors and judges, have a mistaken intuitive sense that there is something especially reliable about eyewitness evidence. Indeed, research has shown that juries tend to overestimate and be particularly influenced by eyewitness testimony (see Wells, 1993 for a review), which, sometimes, has even been proved to be as powerful as a confession (Kassin &

Neumann, 1997). Given this huge impact of eyewitness identification evidence in criminal proceedings, it is concerning how unreliable this evidence can be.

In real-world cases, eyewitness misidentification is the major cause of all wrongful convictions (Scheck, Neufeld, & Dwyer, 2000). To stress numerically the importance of this issue, at the time of writing, evidence from the Innocence Project in the United States (www.innocenceproject.org) indicates that about 75% of the 241 wrongfully convicted people (mostly rape cases because of the possibility of DNA evidence), who have been exonerated since 1989 by post-conviction DNA evidence, were victims of mistaken eyewitness identifications. These wrongfully convicted people served, on average, 12 years in prison and 17 of them served time on death row. Of course the magnitude of the problem can be even bigger than that which I have reported here, given that the frequency of wrongful convictions is not something that can be known precisely. Furthermore, DNA cases are only a small percentage of all crime cases (Wells et al., 1998).

For most of these cases of mistaken imprisonment, the mistaken identification was obtained from lineups, the most commonly used method –but not the only one– for obtaining identification. Sometimes, the innocent suspect was “identified” by more than one eyewitness (Wells et al., 1998).

The Lineup: Current Police Practice

The administration and construction of a lineup varies across different jurisdictions and nations. At its most basic level, a US lineup procedure entails the simultaneous presentation of a photo of a suspect along with the photos of other four, five, or more lineup members who are known to be innocent (hereafter, the *foils*). This type of lineup, where all lineup members are presented simultaneously, is the one that is used with the highest frequency by police in the US (Lindsay & Wells,

1985; Malpass, 2006). Since the 1980s however, sequential lineups, that is, lineups where lineup members are presented one at a time, were introduced by Wells and Lindsay and many organizations have adopted the sequential lineup method (Levi & Lindsay, 2001; Malpass, 2006). Some sequential lineups—dependent on local policies—allow multiple viewing of the set of faces, some allow for only one viewing. Some require that the whole set of faces will be shown to the eyewitness; some require that the procedure stops when the eyewitness makes an identification. In both types of lineups the witness is asked to decide whether the culprit is in the lineup and if so, to indicate which lineup member is the culprit.

In the UK, photo lineups are used seldom as they have been replaced by video lineups. Video lineups have been proved to be fairer than live lineups. In the study of Valentine and Heaton (1998), participants who were given the eyewitness's description of the culprit were asked to guess who the suspect was from a nine-person lineup taken from actual criminal cases. Participants were able to select the suspect only 15% of the time, not significantly above chance level. Video lineups have also been proved to be equally fair for white-European and African-Caribbean suspects (Valentine, Harris, Colom Piera, & Darling, 2003). Also, two recent studies (Valentine, Darling, & Memon, 2007; Darling, Valentine, & Memon, 2008) have found moving images to be slightly superior to still images in reducing foil identifications in target-absent lineups.

There are currently two systems that are used to obtain identification from eyewitnesses using moving images to accord with the guidelines of the code of practice (Code D; Home Office, 2008) enforced by amendments to the 1984 Police and Criminal Evidence Act (hereafter, *PACE*). Where moving images are not possible to be used, still images are the next appropriate option.

Video Identification Parade Electronic Recording (VIPER). UK police have created an enormous face database using people from the general public who volunteer to serve as foils in lineups. Each person is recorded in a short video clip in front of a grey background screen. The clip starts with the volunteer exposing their face and neck looking at the camera for four seconds. Then the volunteer turns to the left exposing the right profile for four seconds and then turns slowly to the right exposing the left profile for another four seconds. Finally, the volunteer turns again to the camera exposing their face for four seconds. To accord with the UK guidelines on lineup size, 8 to 10 similar clips of volunteers –apart from the suspect– are displayed in each lineup procedure. One number on the top left of the screen corresponds to each person, which is used by the witness to identify the person they believe they saw at the crime event.

The system is very efficient time-wise as it can prepare a lineup within half an hour and can be sent to the victim/witness at their home or in hospital. It has also been argued by the West Yorkshire police (who developed this system) that the eyewitnesses/victims might feel more relaxed when they have to face the culprit on the computer screen rather than at the police station. VIPER is the system that is used exclusively by the Scottish Police and half of the police forces in England and Wales.

Profile Matching (PROMAT). The PROMAT system was developed in 1997 for the Manchester Police, and is used by half police forces in England and Wales. In essence, it is very similar to the VIPER system and as such it allows lineups with moving images but the two systems do not use the same face database. PROMAT users collect and share among them their images whereas VIPER is a system that is run nationally; there is a national centre which collects all the images and provides strict checks on their quality before making them available to its users.

Note that both systems provide sequential lineups allowing for two viewings. Both systems are used in a way to accord with the PACE guidelines that state that the images should be moving and that the foils should resemble the suspect (not the eyewitnesses' description of the culprit). For this reason, care should be taken not to select foils that are too similar to the suspect, to an extent that an eyewitness with good memory would not be able to make a positive identification.

Psychological Research

Although the lineup was introduced as a pre-trial safeguard against mistaken identifications, it has proven to be highly unreliable. For example, Valentine and Heaton (1999) showed research participants photos of English police lineups and asked them which of the lineup members they thought that was the suspect. Participants were able to select the suspect 25% of the time in a nine person lineup. If the lineup were perfectly fair (i.e. every lineup member has equal chance of being selected) participants should be able to select the suspect only 11% of the time. Even when properly conducted, lineups elicit 25% to 45% selection rates of the innocent suspect (e.g., Malpass & Devine, 1981). Levi's (1998) meta-review of 47 experiments indicates that an average of 60% of experimental witnesses chooses an innocent foil in a target-absent lineup.

Given these facts, it is very important that we first examine the conditions under which misidentifications are more likely to occur. Then we should focus on how best to design procedures that minimize the danger of misidentification under these specific circumstances. Ongoing research in the area of eyewitness testimony is providing an understanding of why such identification errors occur (see Wells et al., 1998 for a review) by shedding light on the conditions under which mistaken

identifications can be quite high. The literature distinguishes between two different kinds of variables:

Estimator Variables. One category entails factors inherent in the event, the so-called *estimator variables* (Wells, 1978), which include the quality of the viewing conditions, the amount of attention paid by the witness at the time of the event, facial attributes, witness characteristics like gender and age, the time interval between the event and the identification task, and other factors over which the police and the justice system have little control; the actual values of these variables and any possible effects may be estimated but not controlled.

System Variables. The second category entails factors that are under the control of the criminal justice system and are called *system variables*. Some common examples include whether or not witnesses are separated before they have the opportunity to speak with one another and the techniques used by an investigator to elicit a description from a witness. Recent studies have found that a number of system variables are influential during the construction and administration of photo lineups (see Wells et al., 1998 for a review) and from this line of research arose several recommendations on the procedural aspects of eyewitness identification that have been included in the US Department of Justice's guide of best practice for handling eyewitness identification evidence (Technical Working Group on Eyewitness Evidence, 1999, hereafter, the *Guide*). Here I will provide only a brief synopsis of the six most relevant recommendations.

When police investigators prepare to present a photo lineup to the eyewitness, several key factors may influence the likelihood of a misidentification, hence the Guide recommends the following: First, the lineup should be double-blind, that is, neither the eyewitness nor the lineup administrator should be aware of who the

suspect is. Second, instructions should be unbiased, that is, eyewitnesses should be informed that the culprit might or might not be in the lineup and that photographs may not appear exactly as they did on the date of the incident because features such as head and facial hair are often subject to change, although recent research (Charman & Wells, 2007) shows that the changed-appearance instruction increases foil-identification rates without increasing target-identification rates. Third, a confidence statement about the decision should be recorded before any feedback is given. Fourth, the whole procedure should be videotaped for later reference (e.g., for post hoc review of possible suggestions, or for analysing the reaction of the witness to the lineup). Fifth, research indicates that sequential procedures rather than the typical simultaneous ones (in which the eyewitness views all photos at the same time) produce a lower rate of mistaken identifications (in culprit-absent lineups) with little loss in the rate of accurate identifications (in culprit-present lineups). Therefore, use of sequential lineups, in which the eyewitness views one photo at a time and makes an identification decision before viewing the next photo, is required. The superiority of the sequential lineup has been shown in many studies (see Steblay, Dysart, Fulero, & Lindsay, 2001 for a meta-analysis) and it is mainly reflected on the fact that in target-absent lineups, there is a higher probability that the eyewitness will make a correct no-identification decision, resulting to a significant decrease in foil-identification rates compared to simultaneous lineups. However, in target-present lineups identification of the targets is less likely in sequential lineups. Finally, “the suspect should not stand out in the lineup or photospread as being different from the distractors based on the eyewitness’s previous description of the culprit or based on other factors that would draw extra attention to the suspect” (p. 630). This last rule is

the one that is of most concern in this thesis and is explained in more detail in the following paragraph.

Selecting Foils

Wells (1984) showed that when the innocent suspect is the only member of the lineup that fits the eyewitness's description of the culprit, the eyewitness will most likely select this person as he fits the eyewitness's memory of the perpetrator more than any other lineup member, after comparing the photos to each other. This *relative judgment strategy* is efficient for the cases in which the actual culprit is present in the lineup (Lindsay & Wells, 1980) with some exceptions (such as when the foils resemble the suspect too closely, Wells, Rydell, & Seelau, 1993), or when the culprit's appearance has changed from the time of the event to the time of the identification task (Charman & Wells, 2007). However, the relative judgment strategy does not entail a mechanism for rejecting the lineup as culprit-absent when the police have caught the wrong person, increasing the likelihood of a mistaken identification (Wells et al., 1998). The fact that the probability of selecting the suspect is higher when he is the only member of the lineup that matches the eyewitness's description of the culprit, is well established in the literature (e.g., Doob & Kirshenbaum, 1973; Luus & Wells, 1991; Wells, Leippe, & Ostrom, 1979; Wells, Rydell, & Seelau, 1993).

The suspect may stand out because he has some physical characteristics that the other lineup members do not have, or due to different posing, different lighting, background, or simply because he or she more closely fits the description of the perpetrator than the other lineup members do. Figure 1 provides an example of a highly biased, real-world lineup conducted in 1995 in Texas. In this case, the eyewitness had described a black perpetrator; however, all of the foils were Hispanic,

making the suspect stand out. The suspect might as well be standing by himself. Even if the eyewitness had said nothing about the appearance of the perpetrator, if the police had arrested a black man, it would be wrong to stand him in an otherwise all-white, or all-Hispanic in this case, men. Figure 2 provides an example of an unbiased lineup conducted in West Virginia. Here the eyewitness described the culprit having crossed eyes. The special investigator used a computer and forensic art techniques to create crossed eyes in photos of males that they had on file, so that the suspect would not stand out.

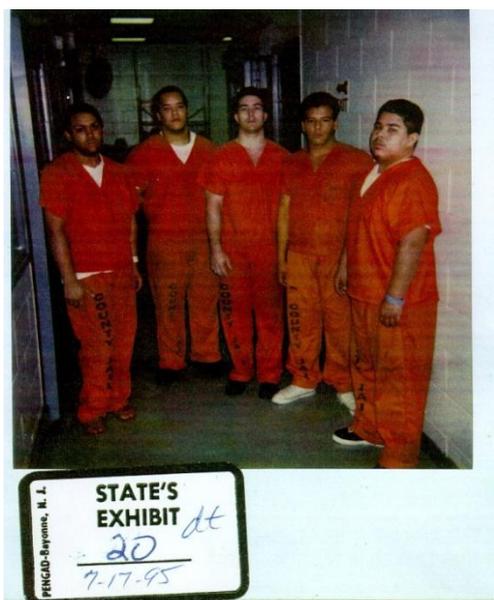


Figure 1. Example of a biased lineup. The image is taken from Gary Wells's homepage on the Iowa State University's website (<http://www.psychology.iastate.edu>).

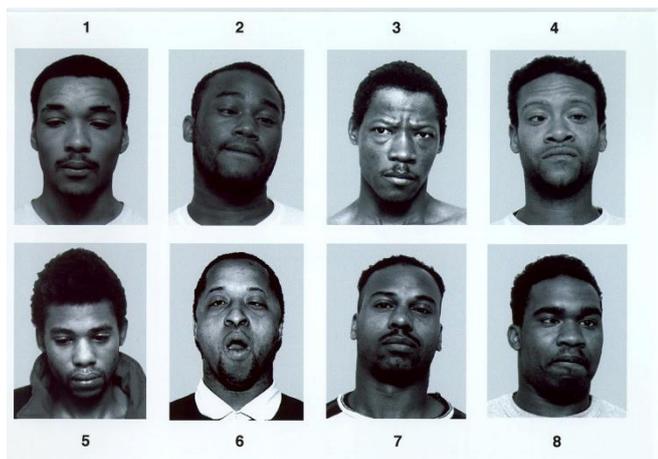


Figure 2. Example of an unbiased lineup. The image is taken from Gary Wells's homepage on the Iowa State University's website (<http://www.psychology.iastate.edu>).

Because eyewitnesses have the tendency to apply relative judgment and pick someone even when the culprit is not in the lineup, unbiased lineups are considered crucial in an effort to reduce this effect (Malpass & Devine, 1981). When eyewitnesses are informed that the culprit might or might not be in the lineup, eyewitnesses are less likely to rely solely on relative judgment strategy and as a result, foil identifications are reduced in target-absent lineups. In target-present lineups though, there is no effect on identifying the culprit (see Steblay, 1997 for a meta-analysis). However, recent studies have revealed a superiority of the simultaneous lineups for the cases where the foils are chosen based on the eyewitness's description rather than according to their resemblance to the suspect (Levi, 2006; Memon & Gabbert, 2003).

The same effect exists when using *blank lineups* (Wells, 1984). A blank lineup consists entirely of foils; in order to test whether the eyewitness is able to resist applying a relative judgment strategy, the suspect is purposely excluded from

the lineup. Reliable eyewitnesses (i.e. those who do not pick a person) are less likely to select mistakenly a foil in a subsequent lineup that does include a suspect.

So, although the selection of foils seems simple at first glance, it is actually very complex and one of the most active issues in the eyewitness identification literature. Ideally, the foils should ensure that every lineup member –including an innocent suspect– has an equal chance of being selected as the culprit. But to what extent should the suspect resemble the foils? Many researchers argue that the foils should not look like the suspect but instead they should be selected based on the eyewitness's description of the culprit (Luus & Wells, 1991). A lineup with foils who very closely resemble the suspect will be fair but it won't be sensitive enough to allow a reliable eyewitness to make a positive identification due to the increased homogeneity among the lineup members. Contrarily, a lineup with foils who match the description of the culprit is still fair but it also creates a propitious heterogeneity among lineup members, which does not interfere with recognition of the culprit (Wells, Rydell, & Seelau, 1993).

The question that immediately arises is how we can construct lineups where both sensitivity and fairness will be at a desirable level. Turtle et al. (1995) suggest an *iterative* strategy. This strategy requires that we initially select photos of people who fit the description of the suspect and not the suspect himself. From this pool of photos we should then select as our first lineup foil the person who resembles the suspect as closely as we want. Then we should put the suspect's photo out of sight and select as the second foil a person that resembles as closely as we want the first foil we just selected. Then we should put the photo of the first foil out of sight and continue following this strategy until we have all the members of the lineup plus one more foil. Then we should exclude from our set of foils the first foil we selected and

use the rest of them to construct a fair and sensitive lineup; all of the foils match the eyewitness's description of the culprit but none of them resembles closely the suspect.

Turtle et al. (2003) note that, in cases that have more than one description because there is more than one eyewitness, and these descriptions vary to an extent that using the same lineup would be unfair, the iterative strategy should be followed to construct a different lineup for each eyewitness based on their description.

Measuring Fairness

Once the lineup has been constructed, there is a widely used method that can be employed to assess systematically the fairness of the lineup: it is the *mock-witness test* (Doob & Kirshenbaum, 1973). During a mock-witness test, independent judges – who have never seen the suspect but are given the eyewitness's description– view the lineup and they are asked which person they think is the suspect. If the lineup is perfectly fair, then each lineup member should be chosen equally often as any other lineup member. For example, if 60 independent judges assess the fairness of a six-person lineup, then every lineup member should be picked as the suspect 10 times. So, knowing the proportion of mock-witnesses who selected the suspect gives us a good idea about how fair the lineup is by comparing it to the expected-by-chance rate.

Once we know the number of mock witnesses who chose the suspect, we can also calculate the *functional size* of the lineup, as another measure of lineup fairness (Wells, Leippe, & Ostrom, 1979) which tests whether the foils that we chose to put alongside the suspect are plausible alternatives to the suspect. Obviously, even if the actual lineup size is 9 but only one of the foils is plausible, there is a higher probability of a mistaken identification if the suspect is innocent and a higher

probability of a correct identification if the suspect is the culprit. The functional size is calculated by dividing the number of mock-witnesses who took part by the number of mock-witnesses who chose the suspect. So, in a six-person lineup, if 30 out of 60 mock-witnesses chose the suspect, the value of the functional size is $60/30 = 2$, which is below the nominal size of the lineup (i. e . the number of lineup members). This indicates an unfair lineup since the suspect was selected at a level above chance ($60 \times 1/6$). Brigham, Ready, and Spier (1990) suggest that the functional size should be greater than half the nominal size. PACE requires 9-person lineups, which means that the functional size of all English lineups should be greater than 4.5. However, recent studies (Brigham et al., 1990, 1998; Wells & Bradfield, 1998) have reported much lower values that go down even to 1.2.

So far, it has become clear that the distinctiveness of the suspect within the lineup increases the likelihood of a mistaken identification. However, very often suspects have distinctive facial features, like scars, birthmarks, tattoos, moles and so on. For these cases, necessary care must be taken so that the suspect (and none of the foils) won't stand out in the lineup and, at the same time, the foils will be plausible alternatives to the suspect.

Current Police Lineup Techniques for Suspects with Distinctive Features

At the time of writing, there is no specific published research that tests the different techniques that are used to construct lineups for suspects with distinctive features. Nonetheless, PACE states that when a suspect has some sort of distinctive facial feature such as a tattoo, facial hair, or an unusual hairstyle, which does not appear on the other lineup members, the distinctive feature should be replicated across lineup members (*replication technique*). Figure 3 shows two examples of how VIPER software is used to replicate digitally a distinctive feature on the face of a foil

(replication). Note that in the case of replication, video lineups are replaced by still, full-face images.



Figure 3. A face (a) before and (b) after the digital alteration to include baldness, a blemish, a moustache, a beard (c) and glasses. The photos are provided by the VIPER National Bureau.

PACE states that, if replication is not possible, the distinctive feature should be concealed on the suspect (*concealment technique*). Figure 4 shows the current techniques that are usually used to conceal a distinctive feature: (a) pixelating the area of the distinctive feature (pixelation), and (b) covering with a black rectangle the area of the distinctive feature. If the distinctive feature is pixelated or blacked out, then the corresponding area on all of the lineup members is also pixelated or blacked out. Replication and concealment serve to ensure that the suspect does not stand out in the lineup. It is at the identification officer's discretion to choose which technique to apply, but PACE requires the decision and the rationale behind it to be recorded. When a culprit's distinctive feature is not reported by the eyewitness but the suspect has a feature deemed by the police to be distinctive, PACE states that concealment should be applied.



Figure 4. Concealment is achieved by either (a) pixelating the area of the distinctive feature, or (b) covering the area of the distinctive feature with a black rectangle. The photos are provided by the VIPER National Bureau.

In the US, an extension of the fit-to-description rule of the *Guide* distinguishes between three different cases:

1. *The eyewitness's description of the culprit does not match the physical characteristics of the suspect or the description is vague.* In this case, the Guide suggests that the foils should have both the features mentioned in the description and the features of the police suspect. If this is not possible, then the selection of foils should be based only on the features of the suspect to prevent him from standing out in the lineup. Attention should be paid to any feature that is mentioned. So, in cases where an eyewitness describes the culprit having, say, a scar, every member of the lineup should have a scar. If the eyewitness reported the scar but the suspect has no scar, then none in the lineup should have a scar. Consistent appearance should by all means be obtained, either by replicating artificially the distinctive feature across lineup members or by concealing it on the face of the suspect.

2. *The eyewitness did not mention a distinctive feature that the suspect has.* In this case replication of the feature across lineup members is not recommended by the Guide, so that recognition memory can come into play, unless, of course, that the suspect stands out to a degree that mock witnesses would be able to select him above chance level. The rationale is that each of the foils will have some unshared features that, if the regions of these distinctive features were isolated and compared to the equivalent regions of the faces of the rest of the foils, would make them stand out. For example, one of the foils may have larger ears or smaller nose than the rest of the foils, which does not necessarily make him stand out.
3. *The eyewitness gives a very detailed description of the culprit.* In the case where very distinctive details are described, the need to construct a lineup is questioned. The police should just apprehend a suspect who fits this description. The lineup serves as a test of memory in cases where the description of the eyewitness is quite vague. In the cases where the description is detailed to an extent that details about the face of the suspect are fully described, a recognition memory task seems unnecessary.

At the time of writing, the only information that I have on what police officers in the US actually do when dealing with suspects with distinctive features comes from the study of Wogalter, Malpass and McQuiston (2004), who provide results of a survey of 220 jurisdictions. The authors developed a 67-item questionnaire to be completed by the most experienced lineup administrator in each jurisdiction, asking them, among other things that are not the scope of the present thesis, what they do when a suspect has distinctive facial markings (e.g., scars or birthmarks) providing them with a list of non mutually exclusive options. According

to their results, 77% try to replicate the marks to every other member of the lineup, 23% try to add similar marks to the other lineup members and 18% try to conceal the area of the markings. Thirty percent answered that they do not do anything with respect to facial marks (percentages sum to more than 100% because some officers reported more than one method).

But do these current techniques used by the police actually prevent suspects with distinctive features from being picked out from a lineup? And are both techniques equally effective? The aim of the present experimental work is to obtain a better approximation to these questions. First, an explanation of the effects of face distinctiveness is to be provided through a review of relevant theoretical models of face recognition.

Models of Face Recognition

In an attempt to explain facial perception and recognition, researchers have traditionally distinguished between two kinds of facial information, namely featural (i.e., isolated features such as nose, eyes, etc.) and configurational (i.e., spatial relations between the features, such as brow area compared to face area, etc.). Researchers adopting a featural hypothesis state that a face is perceived as a sum of its parts (i.e., the individual features) whereas those adopting a configurational hypothesis place greater emphasis on the configuration of the features without neglecting the importance of the individual features. A third hypothesis that has been developed to explain perception and recognition of faces is the holistic hypothesis: in order to remember faces we use both featural and configurational information, hence the face is perceived as a whole. A fourth approach is the norm-based hypothesis which states that both featural and configurational information is represented as deviations from an abstract prototype. This last approach is part of Valentine's

(1991) multidimensional face-space framework that accounts for distinctiveness effects and will be fully discussed later in this section.

Distinctiveness effects. In general, recognition of a face improves as a function of distinctiveness. However, it is important to note that distinctiveness is not an absolute quality. The level of distinctiveness of a given face depends on the observer's experience with several other faces. In other words, the level of distinctiveness does not depend on the face itself, but on its relation to a large number of faces that the observer has encountered and holds in memory. Distinctiveness can depend on unusual features but also on the general structure of the face and can only be judged in relation to a given population (Murdock, 1960).

So, how can we measure the degree of distinctiveness of a face? The distinctiveness of a given face is usually determined by obtaining subjective ratings from experimental participants. Participants are asked to indicate on Likert-type scales the distinctiveness of faces and the resulting mean value assigned to each face represents its level of distinctiveness. Distinctiveness is usually determined by the question "How easy would it be to select this person out of a crowd?", although other questions have also been found in the literature (e.g., How sure are you that this face is "average"?). (Vokey & Read, 1992). Interestingly, Bruce, Burton, and Dench (1994) found a positive correlation between these subjective measures of distinctiveness and a more objective method that they developed to measure the degree of distinctiveness of a face: they measured the degree of deviation of specific facial measures such as width of mouth, and distance between nose and mouth, from a prototype, average face.

In a standard old/new recognition memory task participants are shown a set of previously seen faces and new unseen faces and are asked to identify faces

previously seen. Would we expect that as the level of a face's distinctiveness increases, the probability of the face to be recognized increases as well? The answer is affirmative. It is well established in the literature that distinctive faces, although classified as a face more slowly (Valentine & Bruce, 1986a), are much more likely to be accurately recognized than typical, non distinctive faces. In particular, distinctive faces have been shown to elicit higher hit rates than typical faces when they are used as target faces (Bartlett, Hurry, & Thorley, 1984; Shepherd, Gibling, & Ellis, 1991; Valentine, 1991) and lower false-alarm rates when they are used as foils (Light, Kayra-Stuart, & Hollander, 1979; Valentine, 1991). Shapiro and Penrod's (1986) meta-analysis of face recognition memory studies revealed the same pattern: a higher hit rate and a lower false alarm rate for distinctive faces. Furthermore, typicality (as the opposite of distinctiveness) of faces is one of the most reliable predictors of recognition performance (Deffenbacher, Johanson, Vetter, & O'Toole, 2000; Light, et al., 1979).

However, Vokey and Read (1992) showed that the concept of distinctiveness is much more complex. They used a principle components analysis of subjective ratings of face distinctiveness, likeability, familiarity, memorability and attractiveness to examine how these different factors affect recognition. They found that distinctiveness entails two components: memorability (how easy a face is to remember), which relates positively to hits and negatively to false alarms, and context-free familiarity (the extent to which participants thought they might have seen the face around campus), which relates negatively to false alarms but it does not predict hits.

It seems, then, that the rating of distinctiveness for a given face depends on two factors. The first factor is the context-free, or structurally induced familiarity,

which is a general feeling of familiarity unconnected to a particular context and therefore, typical faces gain more context-free familiarity than distinctive faces. The second factor is the familiarity due to prior exposure. In this case, the feeling of familiarity comes from matching the target face to an item in memory or from perceiving that a face is familiar due to prior exposure. Because encoding and retrieval processes give benefit to distinctive rather than to typical faces, distinctive faces gain more familiarity due to prior exposure than typical faces (Bartlett et al., 1984).

But since the participant is not able to distinguish between the two different types of familiarity, typical faces seem more familiar than distinctive faces because they are more similar to other background faces. Hence a typical face that has not been seen before is more likely to be falsely categorized as previously seen. Similarly, a distinctive face that has not been seen before does not evoke a feeling of familiarity due to its low structurally induced familiarity: participants notice the distinctiveness of the face and assume that if the face was previously seen they would have remembered it. Hence, unseen distinctive faces are easily rejected as new. However, distinctive faces that have been seen before are likely to be recognized as previously seen due to their benefit at encoding (Bartlett et al., 1984). Vokey and Read (1992) argue that participants assess the memorability of a face and they conclude that if it is high enough they would have remembered the face, but since they do not, they assume that they have not seen it. So, as Wixted (1992) noted, this subjective memorability is not a measure of familiarity resulting from matching the target face to faces in memory; it is rather a metacognitive process.

It is interesting to examine how these two feelings of familiarity will apply to a lineup identification task, where the participant is deciding between a target face

and a foil (forced-choice task). Based on Bartlett et al.'s (1984) finding that typical faces evoke more context-free familiarity whereas distinctive target faces have a higher advantage of prior exposure, Busey (2001) speculates that the target face will evoke a higher level of study-induced familiarity combined with a certain level of context-free familiarity. If the foil is very typical though, its context-free familiarity might exceed that of the target face's overall familiarity; hence the foil might be selected over the target face. Busey adds that the selection between two faces where both the target and the foil are distinctive, would be less difficult: in this case, the two faces would evoke equally low levels of context-free familiarity, so the distinctive target face would evoke an overall higher level of familiarity due to the benefit from study.

O'Toole, Deffenbacher, Valentin and Abdi (1994) using digitized pictures, found that small distinctive features were associated with the memorability element of recognition whereas global aspects of the shape of the face were associated with the general familiarity element of recognition. When a face has a small distinctive feature, this will be used as a highly predictive component whereas in the absence of the distinctive feature, more generic information of the face will be used, e.g., the shape of the face.

One explanation of the distinctiveness effects (recognition of distinctive faces is better than recognition of normal faces) was given by the multidimensional face-space framework proposed by Valentine (1991). In this framework, faces are represented as points or vectors in a multidimensional space. A key point in this model is the importance of the large collection of faces that a person has encountered during their life. Within this context, the individual representation of each face and

the degree of similarity among these representations are detrimental to face-recognition performance.

The Multidimensional Face-Space framework. Valentine (1991) uses a multidimensional Euclidean space in which faces are represented. The dimensions represent numerous facial properties, such as hair colour, size of eyes, and shape of face etc., which allow for differentiation among faces. Observers recognize a face by matching the encoding of the target face with the representations of faces stored in memory and located near the target face; the higher the degree of matching, the more probable the recognition. Within this framework, we can distinguish between two different kinds of models: (a) the normed-based model and (b) the purely exemplar based model.

a. The normed-based model states that facial information is encoded as deviations from a facial norm, located at the centre of the face space. A face is encoded as a vector directed from the facial norm to a location in the face space. The vectors of the faces are distributed around the facial norm: there is a higher density of faces around the norm, which decreases as we move away from it. The degree of similarity between two faces is based on their vectors and the angle between them. Computerized caricatures have been traditionally used to manipulate the extent to which a face differs from a facial norm.

b. In contrast, the purely exemplar-based model states that similarity between two faces does not depend on their distance from a facial norm. In particular, the degree of similarity between two faces is distance-based: it depends solely on the distance between the faces.

So, how does this model provide an explanation for the distinctiveness effects? Clearly, the model predicts better performance for faces that remain

distinctive than normal based on the following explanation: A typical, non distinctive face is likely to resemble several other faces in memory, hence it is likely to activate the memorial representations of many known faces rather than just a few.

Conversely, a face with a distinctive feature should resemble few faces in memory and activate only a small number of representations (Valentine, 1991; Valentine & Ferrara, 1991). If so, recognition judgments should be much easier and false identifications should be less likely to occur when faces have permanent distinctive features.

As mentioned earlier, distinctive faces are recognized better than normal faces but normal faces are classified as faces faster than distinctive faces. Valentine (2001) argues that the reason for this is that in order to classify a face we do not need to recognize it (differentiate it among a number of faces); we only need to decide whether it looks like a typical human face. Therefore, typical faces are classified as faces faster because they are closer to the centre of the face-space where there is a higher exemplar density.

An alternative account for the distinctiveness effects was provided by Bartlett et al. (1984). The authors suggested that there is a connection between familiarity and distinctiveness. The basic idea is that repeated exposure to a distinctive face increases familiarity more than repeated exposure to a typical face, hence distinctive faces are easier to remember than typical faces. However, Valentine and Bruce (1986b) showed that familiarity and distinctiveness are not correlated variables despite the fact that memory for faces is affected by both variables.

The Generalized Context Model (GCM). According to the exemplar models of recognition, when participants view a test item, they compare it in memory to the study items. Each of the study items is represented in the cognitive system as an

individual exemplar. The familiarity of a given test item is a function of the similarity of that item to the sum of the exemplars stored in memory. The higher the level of global activation the more familiar the test item seems and therefore the higher the probability that the given item will be judged as “old” by the experimental participants. Nosofsky’s (1986) GCM model is one of the models that implements the summed similarity rule described above and is a representative of the exemplar models within Valentine’s face-space framework described earlier.

A crucial point in Nosofsky’s GCM model as applied to old-new face-recognition experiments is the number of faces which participants have been exposed to during the study phase and which they store in memory. Each of these study faces is represented as a point in the multidimensional face space. The degree to which two faces are similar depends on the distance between these faces in space; the closer they are, the more similar they are. Each time participants are faced with a test face, they compare it to the sum of the faces they have seen during the study phase before they make an old/new decision. Whether a face will be judged as old or new is based on this comparison: the higher the degree of similarity, the higher the degree of familiarity and hence, the higher the probability of participants giving an Old response. Based on this logic, when a new typical test face is compared to old faces, it should evoke a higher feeling of familiarity than a distinctive one and similarly, the false alarm rates should be higher for typical faces than those elicited by distinctive faces. Keeping with the same logic, hit rates for typical faces should evoke a higher feeling of familiarity as they resemble numerous faces in memory and they should increase the probability of being judged as old whereas distinctive faces should have comparably lower hit rates as their summed similarity and hence familiarity is lower as well.

The distinctiveness effect contradicts the summed-similarity decision rule of the exemplar models. As shown, the GCM is unable to account for the high hit rates to distinctive faces. To account for these results, Valentine (1991) proposed that an Identification version of the standard GCM might be more applicable to face recognition. According to this version, distinctive faces get higher hit rates because they are more likely to be encoded in memory than typical faces. The idea is that a distinctive face will have very low summed similarity; hence a larger overall fraction. Although this model can account for the old item distinctiveness, it cannot account for the high false alarm rates.

Nosofsky and Zaki (2003) argue that the GCM cannot account for the effects of distinctiveness because (a) it assumes that all faces have equal self-similarity and (b) it assumes that the summed similarity of distinctive faces will be less than the summed similarity of the typical faces because the distinctive stimuli are less similar to the study exemplars.

The Feature-Contrast Model (FCM). The *feature-contrast model* (Tversky, 1977), instead of using a continuous metric space to define similarity between two faces, uses counts of their common and distinctive features. It states that the more common features two items have, the more similar they are; hence familiarity is higher for these items and so they are more likely to be recognised as previously seen. So this model allows for differing degrees of self-similarity, with an increase in the number of common matching features increasing the measure of self-similarity. Thus, within the FCM framework as applied to face recognition, a face with a distinctive feature will have a higher measure of self-similarity than a face without a distinctive feature.

This idea was also incorporated in Nosofsky and Zaki's *hybrid similarity (HS) model* (2003), which has also been applied to face recognition (Knapp, Nosofsky, & Busey, 2006); a face with a higher self-similarity (due to the presence of a distinctive feature) is more likely to be recognised. The HS model is an extension of the standard generalised context model (GCM, Nosofsky, 1986), taking into account the effects of distinctive features. The HS model is unique in using both the distance between two faces in the multidimensional face-space and the number of common and unique distinctive features that these faces have in measuring similarity between faces. As it already entails a mechanism for distinctive features, the HS model, fully described below, is best suited to make predictions about replicating and concealing distinctive features and is the model that has been used to account for the present data throughout the whole thesis. Other models (e.g., the face space or the WITNESS model described later in this chapter) could have also been used but in these cases, a mechanism for distinctive features would need to be invented.

The Hybrid-Similarity Model (HS). Following the HS model, to make a judgment as to whether a particular face i has been seen before, its global familiarity F_i gets assessed, and this familiarity will determine the probability that the face i will be selected:

$$P(\text{old}|i) = \frac{F_i}{F_i + k} \quad (1)$$

where k is a response-criterion parameter. In the HS model, familiarity is defined as the summed similarity between a test face i and each of the faces seen during the study phase (hereafter, the *exemplars*):

$$F_i = \sum_j s_H(i, j), \quad (2)$$

where $s_H(i, j)$ is the hybrid similarity of face i to exemplar j . The hybrid similarity combines (a) metric measures of similarity between faces as points in a large multidimensional space with (b) feature counting measures of similarity based on counts of the number of shared features and the number of unshared, discrete features. In particular, under the HS, similarity of a test face i to each of the exemplars is given by the equation:

$$s_H(i, j) = C \cdot D \cdot S(i, j), \quad (3)$$

where C ($C > 1$) is a free parameter measuring the *increase* in similarity due to the presence of matching distinctive features, and D ($0 < D < 1$) is a free parameter measuring the *reduction* in similarity due to mismatching distinctive features. Finally, $s(i, j)$ is the *similarity* between any two faces (regardless of the presence of distinctive features) and is a decreasing function of the distance between faces in the metric space. The metric-space similarity of a face i to itself is equal to 1.

So among faces that are equally close to one another in the face-space, faces that have identical distinctive features are the most similar, faces that have similar distinctive features are less similar, and faces that have mismatching distinctive features are the least similar.

Note that the HS model does not entail a choice rule for lineup-identification tasks and so far, it has only been applied to Old/New face-recognition studies.

Throughout the whole thesis, predictions have been based on the assumption that

participants apply a relative judgment strategy while making a lineup identification decision (Wells, 1984; Wells et al., 1998). Participants seem to compare each lineup member to each other and pick the one that is more familiar compared to the rest. Hence we assume that the face with the higher absolute familiarity will be the one that is more likely to be selected as the target. The decision rule is probabilistic, meaning that the higher the difference in familiarity between two faces, the higher the probability that the more familiar face will be selected.

The WITNESS Model. Clark's (2003) WITNESS model is another mathematical model of recognition memory, and it is specific to lineup identification performance.

The WITNESS model is based on Wells' (1984) distinction between relative- and absolute-judgment processes. As mentioned earlier, in a relative judgment strategy, eyewitnesses compare each lineup member to each other whereas in an absolute-judgment strategy eyewitnesses compare each lineup member to their memory of the culprit. Within the WITNESS framework, either judgment is based on the degree of the various matches between the appearance of each lineup member and the eyewitness's memory of the culprit. The "BEST" judgment corresponds to the absolute-judgment strategy and is the match value of the lineup member that most resembles the eyewitness's memory of the perpetrator. The "DIFF" judgment corresponds to the relative-judgment strategy and is the difference in match values between the best and next-best lineup members. The degree of match between the eyewitness's memory of the culprit and the appearance of a lineup member is termed "ecphoric similarity"; a term initially used by Tulving (1981) to refer to the degree to which an item matches the observer's memory of the item.

The WITNESS model states that an eyewitness makes a lineup identification when the weighted sum of these two different kinds of judgment exceed the eyewitness's decision criterion. The probability of an identification is a function of the relation between the various ephoric similarities of the lineup members and the witness's decision criterion. Therefore, lineup manipulations such as change of the target's appearance can be viewed as the causes of a change of one of these two variables. For example, a lineup in which only the suspect has a distinctive feature identical to the culprit's one, may evoke more correct target identifications because of an increase in ephoric similarity when the suspect happens to be the target and more incorrect foil identifications when the suspect is an innocent foil with a similar distinctive feature.

Overview of Thesis

From all the above, it has become apparent that the construction of lineups for suspects with distinctive features may be an important factor in explaining identification performance. Despite the volume of published research and the long debates about the best method to be followed when choosing foils for lineup construction, the effect of distinctive features on lineup identifications as well as the optimal strategies for constructing lineups for these cases has not yet been investigated.

This thesis aims to answer a number of questions. First (Study 1), four empirical crossover recognition-memory experiments examine whether the effects of distinctive features operate during encoding processes, retrieval processes, or during both processes and examine the potential effect of exposure time to the face with a distinctive feature. Second (Study 2), two experiments test whether the results of Study 1 can be generalized to lineup-identification tasks. Replication and

concealment are compared using a face-recognition paradigm in a computer-based experiment with current inmates as stimuli. Third (Study 3), two experiments extend the design of Study 2 to include pixelation in both a face-recognition paradigm and a more ecologically valid, real-world, eyewitness-identification paradigm. Fourth (Study 4), a videotaped simulated crime is applied to test whether it can replicate the results of all previous studies. The hybrid-similarity model of recognition memory has been used to model the results of all the four studies. Finally, I discuss the practical and theoretical implications of these four empirical studies as well as their limitations and directions for further investigation.

Study 1

The Role of Distinctive Features in Face Recognition

Abstract

Eyewitnesses often refer to a distinctive facial feature (e.g., a scar or a tattoo) which may or may not remain intact until the time of the identification task. In other cases, a distinctive feature might not be present or seen during the crime event, yet during the identification task the suspect boasts a distinctive feature. This study reveals that when a face was seen with a distinctive feature during study, was more likely to be recognised when, during test, the distinctive feature was present rather than absent. However, when a face was seen without a distinctive feature at study, the presence of a distinctive feature at test did not affect the probability with which this face was recognised. This pattern of results was apparent across the three different exposure time manipulations (1 second, 5 seconds, and 10 seconds). Theoretical and practical implications are discussed.

Introduction

The appearance of an offender often changes between the time of the criminal act and the time of the identification task due to the addition or removal of distinctive facial features. The police's task under these circumstances is to create a lineup that won't be biased against a suspect due to his distinctive feature and at the same time create a lineup that will facilitate recognition of the offender. This chapter distinguishes between two different scenarios that the police may encounter.

Scenario 1: The Offender Was Seen With a Distinctive Feature

Eyewitnesses often refer to an offender's distinctive feature which may or may not remain intact at the time of the identification task. Absence of the distinctive feature at the time of the identification task may occur either intentionally (e.g., by the offender removing a fake tattoo, a moustache, prosthetics, or piercing) or unintentionally (e.g., blemishes fade away, a bruise or a scar heals). Either way, the eyewitness will be presented with a face which does not match exactly the face that was initially encoded. What is the probability of the eyewitness recognising the offender despite of the loss of the distinctive feature? Is a face that lost its distinctive feature significantly less likely to be identified than a face that retained its distinctive feature? An affirmative answer on this question will have important implications for legal contexts. Even when the offender still has the distinctive feature seen during the criminal act, the police usually conceal the distinctive feature on his face so that he does not stand out in the lineup. This technique is often preferred to replication (in which a distinctive feature is replicated across lineup foils) as a less expensive and more straightforward method. However, there are reasons to believe that concealing the distinctive feature might significantly impair recognition.

First, according to the encoding-specificity hypothesis (Tulving & Thomson, 1973), retrieval is more successful when the contextual information available at encoding is also available at retrieval. Thus the encoding-specificity hypothesis predicts better performance for the cases in which the distinctive feature remains intact, because there is no change in the presentation of the culprit's face between encoding and retrieval. Analogously, for the cases in which the distinctive feature is absent at the time of the identification task, recognition should be impaired.

Second, when a face changes from study to test due to the presence or absence of a distinctive feature, the eyewitness may perceive the identification task to be more difficult. Under these circumstances they may reduce their pre-decisional confidence which will in turn, potentially, influence choice behaviour (Brewer, Weber, & Semmler, 2005).

Third, studies on other aspects of changed appearance are in line with the prediction that faces that lose a distinctive feature between study and test are less likely to be identified. Recognition-memory studies, for instance, show that disguises, changes in pose and facial expression, presence or absence of glasses (Patterson & Baddeley, 1977), changes in visual angle (Bruce, 1982), and the effect of the target's aging (Read, Vokey, & Hammersley, 1990) increase false identification rates. Likewise, lineup-identification studies show that disguises (Cutler, Penrod, & Martens, 1987a, 1987b; Cutler, Penrod, O'Rourke, & Martens, 1986), changes in hair style or facial hair, and the addition or removal of glasses (Read, 1995), result in reduced identification performance. Shapiro and Penrod's (1986) meta-analysis supported these results: When the target's appearance changes from study to test, there is a lower probability that participants will select someone

from the lineup, whereas in cases in which they choose someone, there is a higher probability of a mistaken identification.

The crossover-recognition studies presented in this paper build upon this work by investigating whether the effect of distinctive features occurs during the encoding process, during the retrieval process, or during both processes. Also the present study can be applied to other methods of eliciting identifications, like show-ups, deck identifications and street identifications, where there is only one person to be identified. A signal detection analysis is also conducted in order to control for participants' response biases and the discriminability of the faces. Finally, this study uses distinctive features that have not been investigated in previous research (e.g., moles, scars, and bruises).

Scenario 2: The Offender Was Seen without a Distinctive Feature

There are often cases in which the offender was seen during the criminal act without a distinctive feature, yet he possesses one at the time of the identification task (e.g., the appearance of a mole, a scar, or blemishes). In such cases, the police either replicate the distinctive feature across lineup members or conceal it on the face of the suspect. However, choosing one technique over the other might have important consequences to recognition accuracy. Are faces that remain without distinctive features more likely to be recognised than faces that gain a distinctive feature at the time of the identification task?

For the reasons mentioned in Scenario 1, the encoding specificity hypothesis predicts better performance for faces that remain without distinctive features. If the present study supports this hypothesis, it means that police officers should consider concealing a distinctive feature instead of replicating it for the cases in which the offender did not have a distinctive feature at the time of the criminal act.

Note that there are also cases where the police have caught a suspect who boasts a distinctive feature, yet it is impossible to know whether this feature was present at the time of the criminal act (e.g., the eyewitness gave a vague description, the eyewitness encoded the feature but did not report it, or the eyewitness did not have a view of the specific area of the distinctive feature). In such cases the question that arises is: Are faces that retain their distinctive feature more likely to be recognised than faces that gained a distinctive feature?

To sum up, we predict that faces that retain their distinctive feature will be more likely to be recognised than faces that lose a distinctive feature and faces that remain without distinctive features will be more likely to be recognised than faces that gain a distinctive feature. Furthermore, we predict that faces that remain distinctive will be the most likely to be recognised whereas faces that lose a distinctive feature will be the least likely to be recognised. The rationale is based on the basic principle of the *novel popout effect*, described as a phenomenon according to which unexpected stimuli are more likely to capture attention and are more likely to be encoded in memory than expected stimuli (Johnson, Hawley, Plewe, Elliot, & Dewitt, 1990). The novel popout effect is often cited in reviews of eyewitness memory literature as a factor that impairs eyewitness memory when the unexpected stimulus is encoded in memory at study yet is not present at test (e.g., Brewer, et al., 2005). Put simply, if the focus of the eyewitness's attention is drawn to an unusual feature like a scar, they are less likely to encode other facial features and this is detrimental to recognition in the absence of the scar at test. Conversely, in cases in which a distinctive feature remains intact between study and test (e.g., a birthmark), the novel popout effect serves to enhance memory for the face (presumably because people can recognise the distinctive feature, not the rest of the face). This is because

the distinctive feature may attract more attention during encoding creating a stronger trace in memory; hence it will be more likely to be recognised when it remains intact at test. Therefore, one would expect that memory for faces with distinctive features would be stronger than memory for faces without distinctive features. With respect to retrieval, memories of faces with distinctive features should be easier to match against test faces than memories of non distinctive faces. Equally, a new face with a new distinctive feature would be more likely to be rejected as old compared to a new face without a distinctive feature. The danger though is in the cases where a distinctive feature creates a strong memory trace at study but then appears on another face during test. Then we would predict increased false-alarm rates. For example, if the police arrest an innocent suspect with a distinctive feature that matches the perpetrator on the basis of eyewitness testimony, then that innocent suspect is more likely to be erroneously identified in a subsequent lineup task due to the matching distinctive feature.

The Effect of Exposure Time

A second aim of the present paper is to investigate whether the effect of distinctive features mediates the relationship between exposure time and accuracy. How will participants make use of the extra exposure time at encoding? Will they spend more time encoding the distinctive feature or their attention will shift to other facial features detrimental to recognition?

Is there a positive relationship between exposure time to a face and recognition performance? Although intuitively we might think that the longer the exposure to a culprit is, the more reliable the eyewitness's identification decision will be, real crime stories suggest that often this is not the case. There are many cases of miscarriages of justice resulting after an erroneous identification despite the fact that

the eyewitness had spent long time with the culprit, or had even interacted with them (Memon, Hope, & Bull, 2003).

The role of exposure duration in identification performance has been largely ignored in the eyewitness literature and the findings from the existing studies do not give consistent results, sometimes revealing higher hit rates and lower false-alarm rates as exposure time increases (Laughery, Alexander, & Lane, 1971) and sometimes revealing an accuracy improvement which becomes smaller as exposure time increases (Ellis, Davies, & Shepherd, 1977). In Shapiro and Penrod's (1986) meta-analysis, hit rates increased with longer exposure times but, unexpectedly, false-alarm rates also increased.

Findings from the real-world lineups seem to support the intuitive assumption that longer exposures lead to higher target-identification rates (Valentine, Pickering, & Darling, 2003), but as Brewer et al. (2005) mention, such interpretation of this pattern might not be explicit because there is no way to know the proportion of suspects who are the actual offenders. Therefore the target "identifications" may include both correct and wrongful identifications.

A study by Memon et al. (2003) revealed significantly higher hit rates and lower false alarm rates for the condition of the longer exposure (45 seconds) compared to that of the shorter exposure (12 seconds). However, the unusually high hit rates at the long exposure condition (.95), and the unusually high false-alarm rates at the short exposure condition (.90) led Brewer et al. (2005) to think that their results might have reflected the characteristics of the stimulus and/or the lineup.

In an unpublished study of Vokey, Weir and Read (1988, cited in Read, Vokey, & Hammersley, 1990), increased exposure duration (from 3 to 8 sec.) led to increased identification performance for the cases where identical pictures of faces

were used; the opposite was true when non identical photos were used. Increased exposure duration had a negative effect on identification accuracy. Therefore, we may predict that, in the cases where the faces do not change format between study and test, increased exposure time will increase recognition accuracy. When faces change format between study and test, recognition accuracy won't benefit from increased exposure time. Specifically, if participants use the extra time processing the distinctive feature (due to the novel popout effect, mentioned previously), recognition performance will be increased if the distinctive feature remains the same between study and test but not if it is lost at the time of the recognition task. In sum, there are reasons to believe that the beneficial effect of exposure time might disappear when faces have distinctive features that are absent during test.

Overview of Experiments

In three experiments we applied a crossover recognition-memory paradigm (Deffenbacher et al., 2000). During the study phase, participants viewed a series of 32 faces taken from a college year book. Half of the faces were presented unaltered, hereafter *normal*, and half were presented as *distinctive* after they were altered with Adobe Photoshop CS2 to include a distinctive feature (either a mole, facial hair, a scar, or a bruise). During the test phase, participants saw the *old* faces from the study phase together with an equal number of *new* faces (half of which were normal and half of which were distinctive). Participants were asked to decide for each face whether it had been seen during the study phase. Of the faces presented without a distinctive feature in the study phase, half remained in the same format and half had distinctive features added to them. Of the faces presented with a distinctive feature in the study phase, half remained in the same format in the test phase and half had distinctive features removed from them. So, the design was a 2 (study format:

normal, distinctive) x 2 (test format: normal, distinctive) within-participants factorial design. Exposure time was 5 seconds for Experiment 1, 1 second for Experiment 2, and 15 seconds for Experiment 3.

Experiment 1

Method

Participants. Fifty-five undergraduate psychology students from Warwick University ($M = 21$ years, $SD = 5$, 76.36% female) participated for course credit or for £3 payment.

Apparatus and Materials. The stimuli were presented on a 15.4 TFT LCD monitor with a resolution of 1280 x 800 pixels over a refresh rate of 60 Hz. The controlling and monitoring of the stimulus presentation and response registration was served by a personal computer (PC) programmed with E-Prime (Schneider, Eschman, & Zuccolotto, 2002).

Stimuli. The stimuli were developed especially for this experiment using existing face stimuli from Jones et al.'s (2006) materials. These *original* stimuli were 140 gray-scale bitmap (225 pixels height x 169 pixels width) photographs of faces of European males, none of whom wore glasses or had facial hair. A more detailed description of the original stimuli can be found in Jones et al. (2006).

Eighty of the original faces were randomly selected and distinctive features were digitally added to them using Photoshop, so that the final set consisted of 160 faces; eighty normal (20 for each type of distinctive feature) and 80 distinctive. Figure 1 provides an example of eight stimuli. All photographs were 10.8 cm height x 7.9 cm width, had neutral backdrops, and were presented on a white background. Prior to conducting the experiment, all stimuli had been rated by 40 independent judges with respect to their distinctiveness, pleasantness and degree of arousal

elicited by each face. The results revealed no outliers on the attractiveness scale. On the other two scales, we removed the 16 faces, four for each type of distinctive feature, that scored too highly when presented in their normal, unaltered form, leaving a set of 64 faces that was finally used in Experiment 1.



Figure 1. Examples of normal faces (top row) and the same faces after the digital addition of a mole, facial hair, a bruise and a scar (bottom row, from left to right) as presented in Experiments 1 and 2.

Procedure. Before the study phase, participants viewed 8 example faces, taken from the initial 80-face set to familiarize themselves with the type of faces they were going to be viewing and rating. Four of these faces were normal and each of the remaining four had one of the different four types of distinctive features. These example faces were presented in random order and were not included in the actual experiment.

In the study phase, participants were informed that they would be shown a series of 32 faces, one at a time for 5 s each and subsequently they would be tested

on their memory of these faces. Participants were asked to examine each face carefully and to make two ratings (one for distinctiveness and one for emotional arousal) for each face on 9-point Likert scales. For distinctiveness, a score of 1 indicated *not at all* and a score of 9 indicated *very*. To measure the emotional arousal that participants felt while viewing each face, the corresponding scale of the Self-Assessment Manikin (SAM) was used (Bradley & Lang, 1994) and participants followed the official SAM instructions. The order of the two scales was random for each face. Participants were informed that there were no right or wrong answers, and that they would simply have to respond as honestly as they could by clicking with the mouse on the corresponding button on a scale from 1 to 9. Each trial started with the appearance of a fixation cross centred on the screen for 300 ms. Next a face stimulus appeared for 5 s centred on the screen. The 32 faces were presented in random order.

In the test phase, which followed a 5-minute filler task of solving anagrams, participants completed a recognition-memory test. They were told that they would view another series of faces, some of which had been previously seen in the experiment and some of which would be new. Participants were also instructed that a person previously seen might have a different appearance at test and that their task was to recognize the person previously seen, not the exact photograph. They were asked to give their response by clicking with the mouse on a “Yes” button if the face was previously seen or by clicking on a “No” button if the face had not been previously seen. Participants had to select one of the two buttons; they did not have the option of not responding. Participants viewed the 64 faces, one at a time in random order. On each trial a face was presented until the participant responded and the next face appeared 500 ms after the response. Each response was coded as a *hit*, *miss*, *false alarm (FA)*, or *correct rejection*, based on Table 1. No feedback was

provided during the test phase. The duration of the experiment was approximately 25 minutes.

Table 1.

The Four Possible Responses to Each Face Stimulus.

Face	Response	
	"Yes"	"No"
Old	Miss	Hit
New	Correct Rejection	False Alarm

Results

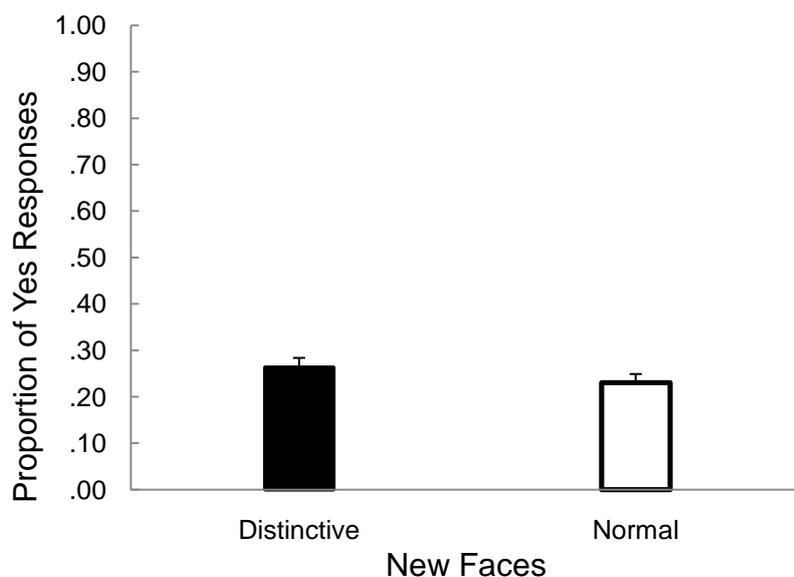
Manipulation Check. The distinctive faces were given the highest ratings on both the distinctiveness and the arousal dimensions by the experimental participants. Mean distinctiveness ratings were significantly higher for distinctive faces ($M = 4.67$) than for normal faces ($M = 3.27$) [$t(54) = 8.985, p < .001, r = .37$]. Similarly, arousal ratings were significantly higher for distinctive faces ($M = 4.93$) than for normal faces ($M = 3.45$) [$t(54) = 10.548, p < .001, r = .36$].

Analysis of FAs to New Faces. Figure 2a shows the proportions of "Yes" responses to new faces. New faces without distinctive features and new faces with distinctive features were equally likely to be rejected as previously seen, $t(54) = 1.667, p = .101$.

Analysis of Hits to Old Faces. A 2 (study format: normal, distinctive) x 2 (test format: normal, distinctive) within-participants ANOVA conducted on hits ("Yes" responses to old faces) yielded the results we expected (Figure 2b). There was a significant interaction between study and test showing that faces that did not change format between study and test were significantly more likely to be correctly identified compared to faces that changed format between study and test, $F(1,54) =$

12.567, $p = .001$, $MSE = .333$, $r = .43$. The ANOVA did not reveal a main effect of study, $F(1,54) = .713$, $p = .402$, $MSE = .016$. Finally, there was a main effect of test, $F(1,54) = 6.037$, $p = .017$, $MSE = .170$, $r = .32$; hit rates were higher for faces that were tested distinctive than for faces that were tested normal. Pairwise comparisons showed that retaining a distinctive feature was significantly better than removing or adding a distinctive feature. No other pairwise comparisons were significant. So, faces that remained without distinctive features were equally likely to be recognized as faces that gained a distinctive feature.

(a)



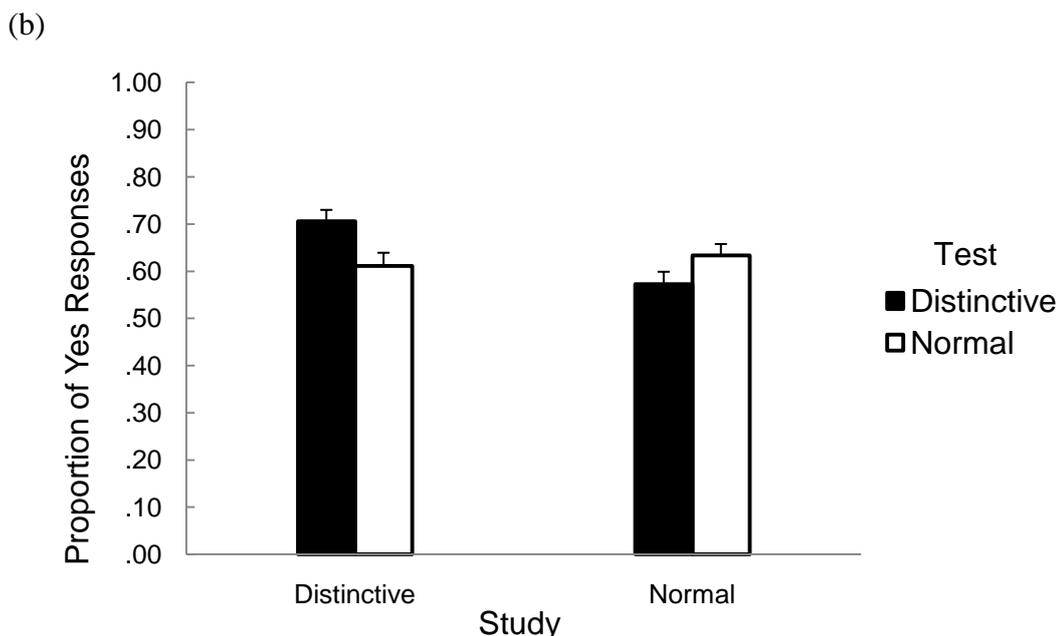


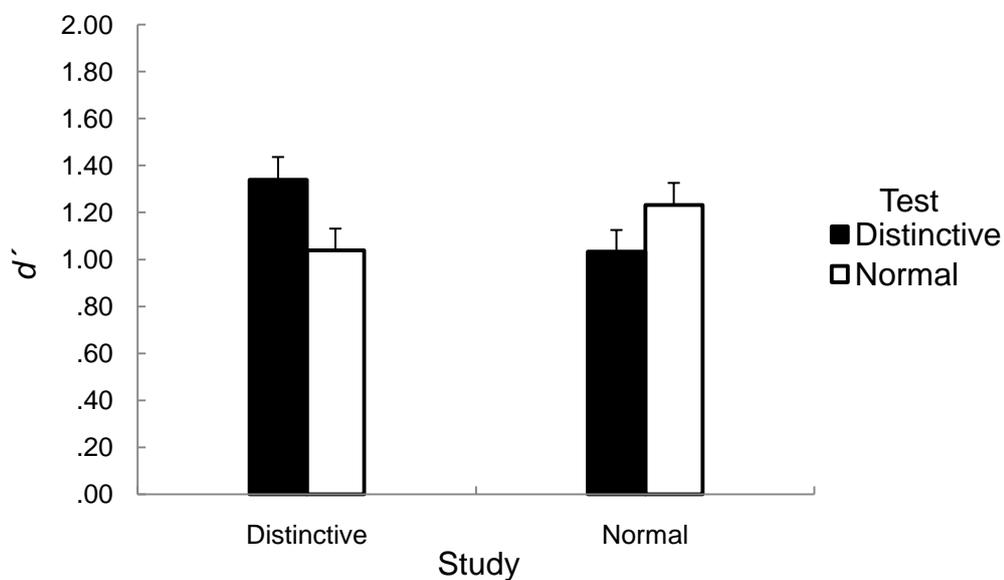
Figure 2. Mean proportions of Yes responses to (a) new faces and to (b) old faces in Experiment 1. Error bars represent the standard error of the mean.

Signal Detection Analysis. Furthermore, in order to estimate how well separated the distribution of familiarity for new and old faces was (d' prime), independently of participants' bias to respond (C criterion), we conducted a signal detection analysis. To avoid infinite values of d' , hit rates and false-alarm rates were adjusted by adding .5 to the number of old responses and dividing by the number of responses +1 (Snodgrass & Corwin, 1988). The non-parametric tests for sensitivity and criterion (A' and B'' respectively) were also calculated and revealed the same qualitative pattern.

A 2 x 2 within-participants ANOVA on d' prime (Figure 3a) revealed a significant interaction between study and test, showing that participants found it significantly easier to discriminate faces that remained in the same format between study and test compared to faces that changed format between study and test, $F(1,54) = 13.770, p < .001, MSE = 3.440, r = .45$. There was no main effect of study, $F(1,54) = .717, p = .401, MSE = .145$, or test, $F(1,54) = .488, p = .488, MSE = .177$. Pairwise

comparisons showed that faces that remained distinctive were significantly easier to discriminate compared to faces that gained or lost a distinctive feature. No other pairwise comparisons were significant.

(a)



(b)

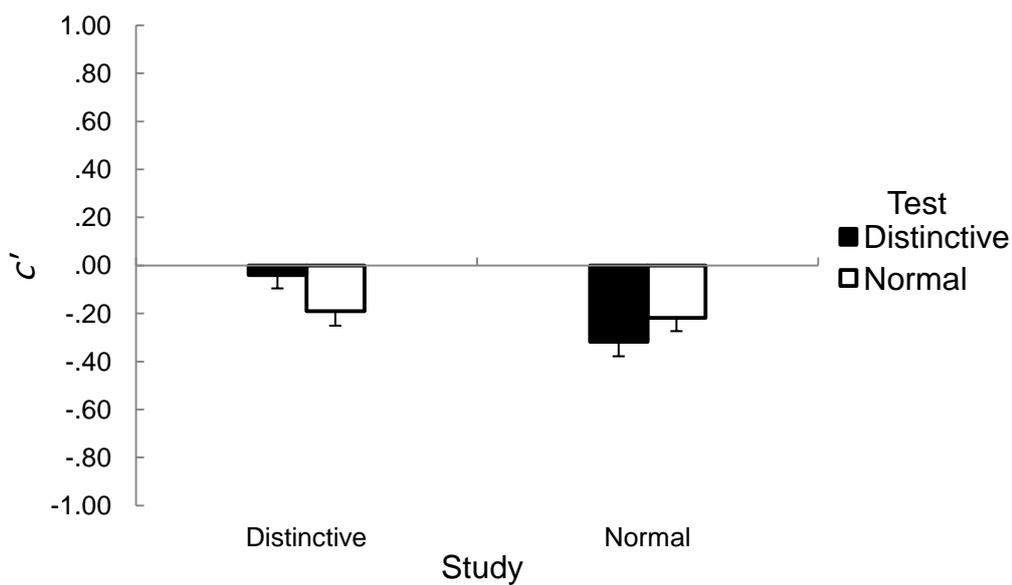


Figure 3. (a) d' and (b) C as a function of study and test format in Experiment 1.

Error bars represent the standard error of the mean.

Finally, in order to detect the tendency of the faces to evoke “old” versus “new” responses, the C criterion values were also calculated and again a 2 x 2 within-participants ANOVA was conducted (Figure 3b). There was a significant interaction between study and test, $F(1,54) = 13.770, p < .001, MSE = .860, r = .45$. When faces changed format, participants became more conservative in their responses whereas when faces remained the same between study and test participants were closer to being ideal observers (C values closer to zero). There was no main effect of study showing that the format of the faces at study, did not affect participants’ bias to respond “old” or “new”, $F(1,54) = .717, p = .401, MSE = .036$. Finally, there was a main effect of test showing that the C values for faces that were tested distinctive were significantly lower than for faces that were tested normal, $F(1,54) = 6.631, p = .013, MSE = 1.283, r = .33$. Pairwise comparisons showed that participants were significantly more conservative for faces that gained or lost a distinctive feature than for faces that retained their distinctive feature. No other pairwise comparisons were significant.

Our results supported our initial hypothesis that faces that remained the same would be more likely to be recognised than faces that changed between study and test. However, this interaction was not symmetrical. From the faces that remained the same, only faces with distinctive features were significantly more likely to be recognised than faces that either lost or gained a distinctive feature. The signal detection analysis showed that these results were caused by a bias that was observed in participants’ responses: participants were responding with a significantly more conservative manner to faces that lost or gained a distinctive feature compared to faces that retained their distinctive feature.

Experiment 2

Method

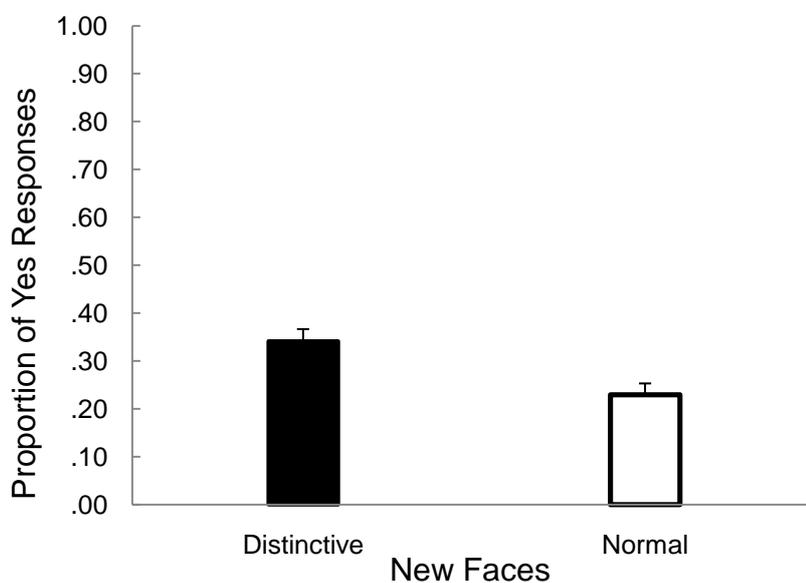
Participants. Thirty-three University students ($M = 27.17$ years, $SD = 8.52$, 52.73% female) participated for course credit or for £3 payment.

Stimuli. The stimuli were taken from Experiment 1.

Design and Procedure. The design and procedure were identical to those of Experiment 1 for one modification: The exposure time for each face stimulus during the study phase was 1 second instead of 5 seconds.

Results

(a)



(b)

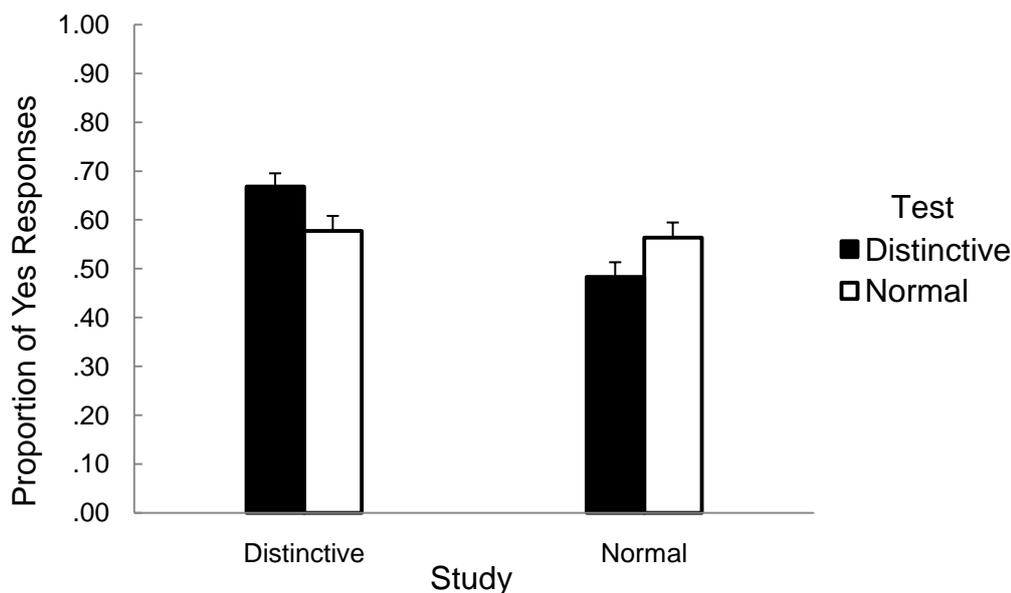


Figure 4. Mean proportions of Yes responses to (a) new faces and to (b) old faces in Experiment 2. Error bars represent the standard error of the mean.

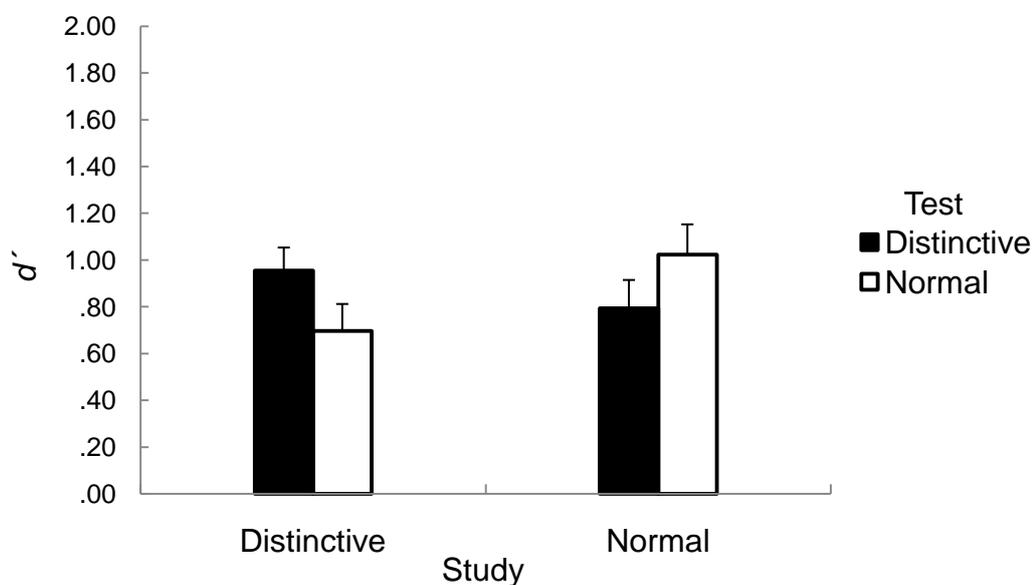
Analysis of FAs to New Faces. Figure 4a shows the proportions of "Yes" responses to new faces. There was a significant difference between distinctive and normal faces with the distinctive faces eliciting a significantly higher FA rate than the normal faces, $t(32) = 3.950$, $p < .001$, $r = .57$. This difference was not observed in Experiment 1.

Analysis of Hits to Old Faces. The results of a 2 x 2 ANOVA conducted on the hits (Figure 4b) replicated the results of Experiment 1. There was a significant interaction between study and test, $F(1,32) = 10.081$, $p = .003$, $MSE = .243$, $r = .49$. The ANOVA did not reveal a main effect of study, $F(1,32) = .039$, $p = .845$, $MSE = .001$. Finally, there was a main effect of test, $F(1,32) = 15.959$, $p < .001$, $MSE = .326$, $r = .58$. Pairwise comparisons showed that faces that remained distinctive were significantly more likely to be recognised compared to faces that lost a distinctive feature, replicating the results of Experiment 1 but not compared to faces that gained

a distinctive feature. Also, contrary to Experiment 1, faces that remained distinctive were significantly more likely to be recognised than faces that remained normal. No other pairwise comparisons were significant in Experiment 2.

Signal Detection Analysis. A 2 x 2 within-participants ANOVA on d' prime (Figure 5a) revealed a significant interaction between study and test, $F(1,32) = 9.888$, $p = .004$, $MSE = 1.967$, $r = .49$. There was no main effect of study, $F(1,32) = .042$, $p = .840$, $MSE = .007$, or test, $F(1,32) = .568$, $p = .456$, $MSE = .224$. These results replicate the results of Experiment 1. However, pairwise comparisons did not reveal any significant differences between any of the face categories.

(a)



(b)

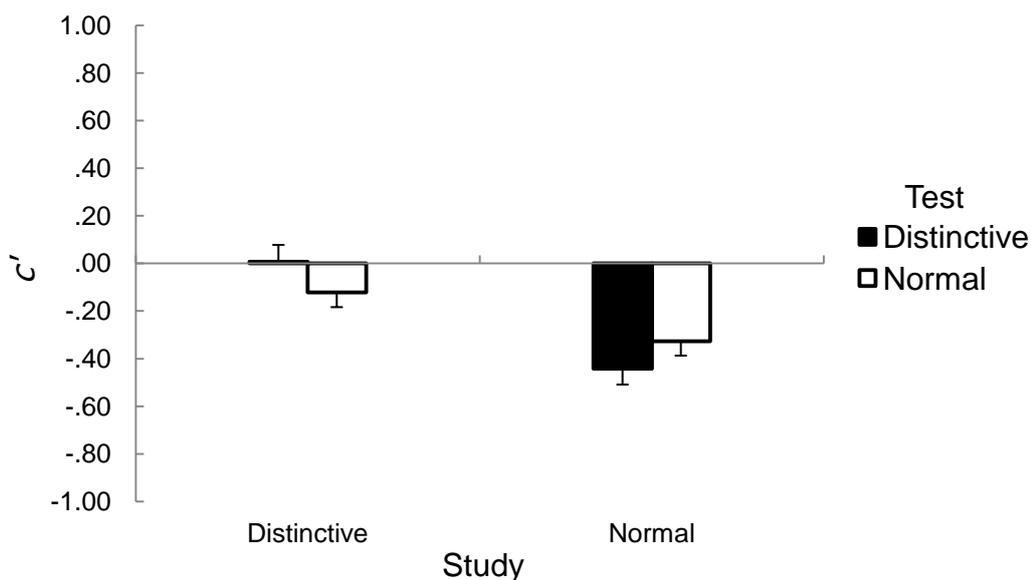


Figure 5. (a) d' and (b) C as a function of study and test format in Experiment 2.

Error bars represent the standard error of the mean.

The results of a 2 x 2 within-participants ANOVA (Figure 5b) conducted on the C values replicated the results of Experiment 1. There was a significant interaction between study and test, $F(1,32) = 9.888$, $p = .004$, $MSE = .492$, $r = .49$, with participants becoming more conservative when faces changed between study and test. Again, there was no main effect of study, $F(1,32) = .042$, $p = .840$, $MSE = .002$. Finally, there was a main effect of test showing that participants were significantly more conservative for faces that were tested as distinctive than for faces that were tested as normal, $F(1,32) = 21.847$, $p < .001$, $MSE = 3.549$, $r = .64$. Pairwise comparisons showed that participants were closer to being ideal observers for faces that remained distinctive than for faces that lost a distinctive feature, gained a distinctive feature or remained normal, all of which rendered more conservative responses.

To summarize, under the 1-second exposure, the same pattern of results was revealed. Faces that remained the same were significantly more likely to be recognised than faces that changed format between study and test but simple-effects analysis revealed that the interaction was not symmetrical. Faces that retained their distinctive feature were significantly more likely to be recognised compared to faces that lost or gained a distinctive feature but this recognition advantage was not observed for faces that remained without distinctive features. The only difference compared to Experiment 1 was that the analysis of the false-alarm rates for the 1 second-exposure condition, revealed a significant difference between the distinctive and normal faces with the distinctive faces eliciting a higher false-alarm rate than the normal faces.

Experiment 3

Method

Participants. Thirty-one University students ($M = 24$ years, $SD = 7.52$, 60% female) participated for course credit or for £3 payment.

Stimuli. The stimuli were taken from Experiments 1 and 2.

Design and Procedure. The design and procedure were identical to those of Experiment 2 for one modification: The exposure time for each face stimulus during the study phase was 10 seconds instead of 1 second.

Results

Analysis of FAs to New Faces. Figure 6a shows the proportions of Yes responses to new faces. There was no significant difference between distinctive and normal faces, $t(30) = .715$, $p = .480$. This result is in line with Experiment 1 (5 seconds) but not Experiment 2 (1 second).

Analysis of Hits to Old Faces. A 2 x 2 ANOVA conducted on the hits replicated the results of Experiment 1 (Figure 6b). There was a significant interaction between study and test showing that faces that did not change format between study and test were significantly more likely to be correctly identified compared to faces that changed format between study and test, $F(1,30) = 40.076$, $p < .001$, $MSE = .703$, $r = .76$. The ANOVA did not reveal a main effect of study, $F(1,30) = .369$, $p = .548$, $MSE = .006$. Finally, unlike in Experiments 1 and 2, there was no main effect of test, $F(1,30) = 2.918$, $p = .098$, $MSE = .067$; hit rates were equally high for faces that were tested distinctive and for faces that were tested normal. Pairwise comparisons showed that, as in Experiments 1 and 2, faces that remained distinctive were significantly more likely to be recognised than faces that lost a distinctive feature and, as in Experiment 1, faces that remained distinctive were significantly more likely to be recognised than faces that gained a distinctive feature. Contrary to Experiment 2 and in line with Experiment 1, faces that remained distinctive were significantly more likely to be recognised than faces that gained a distinctive feature. Also, faces that remained normal were significantly more likely to be recognized than faces that lost a distinctive feature. No other pairwise comparisons were significant.

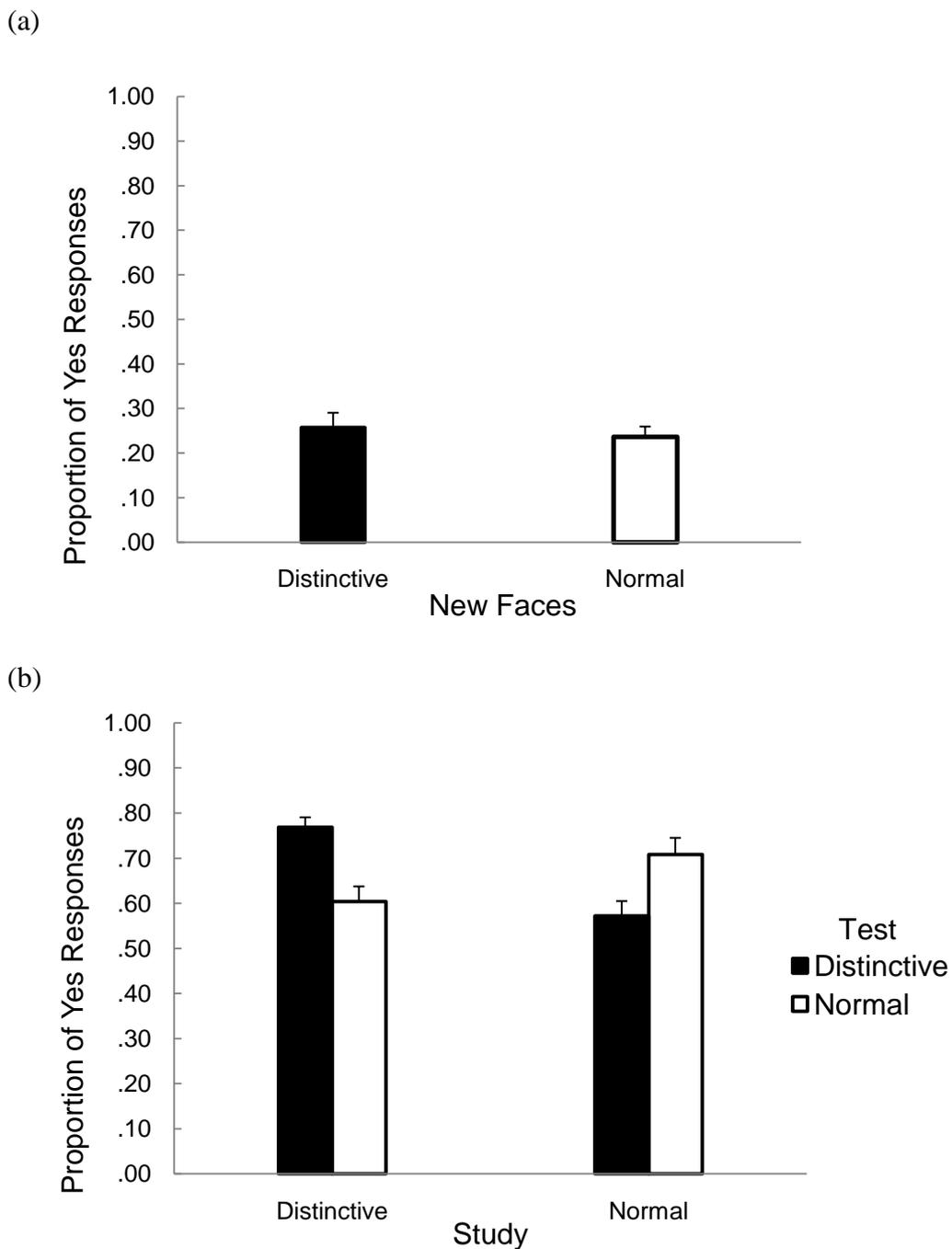


Figure 6. Mean proportions of Yes responses to (a) new faces and to (b) old faces in Experiment 3. Error bars represent the standard error of the mean.

Signal Detection Analysis. The results of the ANOVA on the raw data are supported by the signal detection analysis. A 2 x 2 within-participants ANOVA on d' prime (Figure 7a) revealed a significant interaction between study and test, $F(1,30) =$

37.505, $p < .001$, $MSE = 6.860$, $r = .75$. There was no main effect of study, $F(1,30) = .094$, $p = .761$, $MSE = .007$. Finally, there was no main effect of test, $F(1,30) = .862$, $p = .360$, $MSE = .353$. These results replicate the results of Experiments 1 and 2. However, pairwise comparisons showed that, unlike in Experiment 2, faces that remained distinctive were easier to discriminate than faces that gained or lost a distinctive feature. Also, faces that remained normal were significantly easier to discriminate than faces that lost a distinctive feature. No other pairwise comparisons were significant.

The C criterion values were also calculated and again a 2 x 2 within-participants ANOVA was conducted (Figure 7b). There was a significant interaction between study and test, $F(1,30) = 37.505$, $p < .001$, $MSE = 1.715$, $r = .75$. As in Experiments 1 and 2, when faces changed format, participants became more conservative in their responses whereas when faces remained the same between study and test participants were closer to being ideal observers. Again, there was no main effect of study, $F(1,30) = .094$, $p = .761$, $MSE = .004$. Finally, unlike Experiments 1 and 2, there was no main effect of test, $F(1,30) = .629$, $p = .434$, $MSE = .118$. Pairwise comparisons showed that participants were significantly more conservative for faces that gained a distinctive feature than for faces that remained distinctive. As in Experiment 1 and 2, for faces that remained distinctive, participants were closer to being ideal observers than for faces that lost a distinctive feature for which they were more conservative. Finally, participants were more conservative for faces that lost a distinctive feature than for faces that remained normal. No other pairwise comparisons were significant.

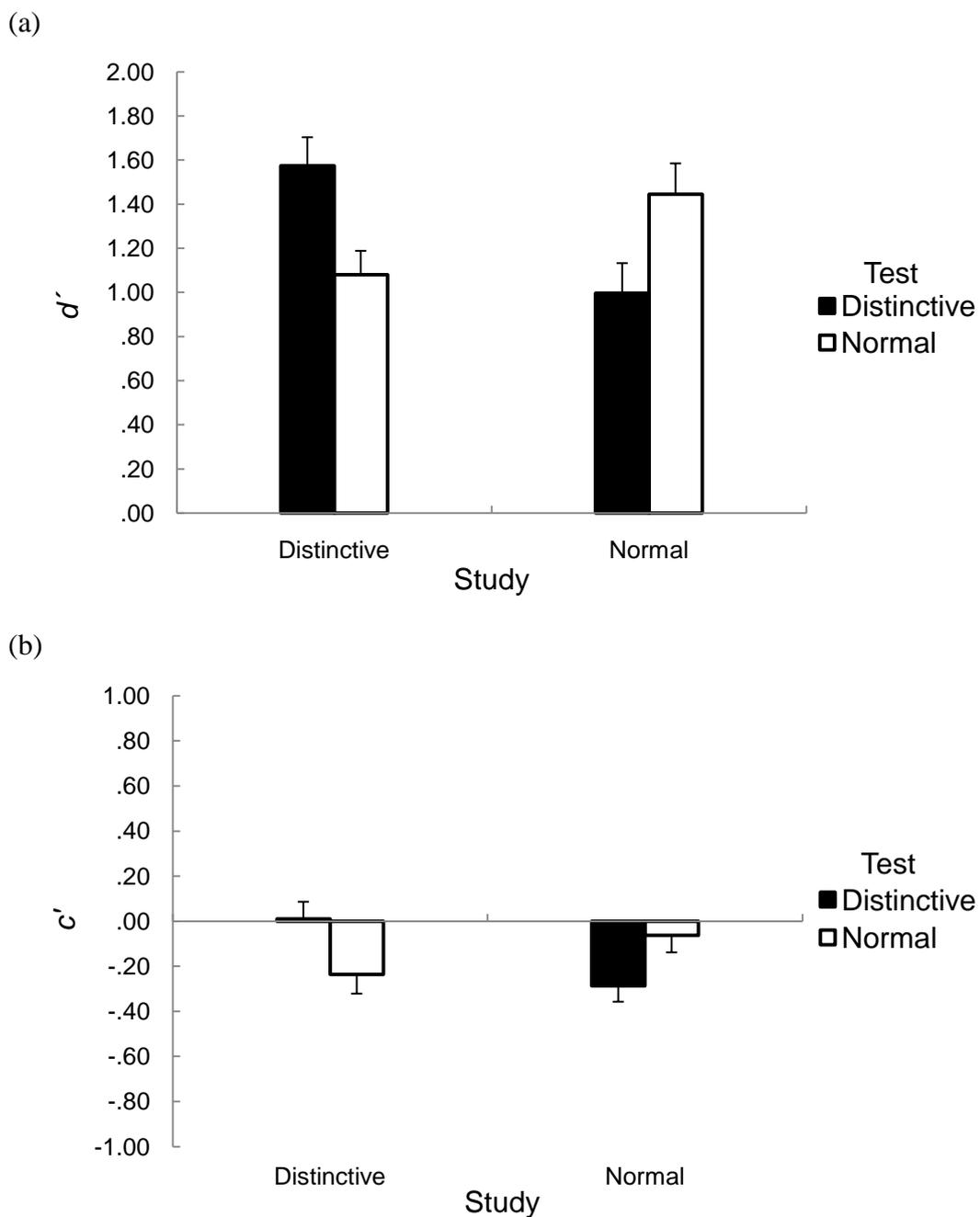


Figure 7. (a) d' and (b) C as a function of study and test format in Experiment 3.

Error bars represent the standard error of the mean.

To summarize, as in the 5-second-exposure experiment, there was no significant difference in false-alarm rates between the normal and the distinctive faces. Concerning the recognition of previously seen faces, the difference in this experiment is that faces that lost a distinctive feature were significantly less likely to

be recognized than faces that remained without distinctive features. Speculations about the reasons of the differences observed among the three experiments as well as the implications of these findings are discussed later.

Modelling. The HS model, as described in the introduction, was applied to data from Experiments 1-3. Because of the random allocation of faces in this experiment, we modelled the similarity between any two faces with the average pairwise similarity, s . For the purposes of this particular experimental design, we introduced two more parameters: M ($0 < M < 1$) is a free parameter measuring the reduction in similarity when a face without a distinctive feature is compared to a face with a distinctive feature, and B ($1 < B < C$) is a free parameter measuring the increase in similarity due to shared distinctive features that are not identical but they are of the same type. So among faces that are equally close to one another in the face-space, pairs of faces that have identical distinctive features are most similar, pairs of faces that have the same type of distinctive feature are less similar, pairs of faces that have mismatching distinctive features are the less similar still, and pairs of faces in which only one face has a distinctive feature are least similar. Table 2 shows the similarity of test faces to the 32 study faces. For example, the familiarity of an old distinctive test face i is given by its similarity to the study faces and is

$$C + (3 B s) + (12 D s) + (16 M). \quad (4)$$

The first term, C , measures the old distinctive face's self-similarity; the second term, $3 B s$, measures the summed similarity of the old distinctive face to the 3 distinctive old faces with similar distinctive features, the third term, $12 D s$, measures the summed similarity of the old distinctive face to the 12 old distinctive faces with

dissimilar distinctive features; and the final term, $16 M$, measures the summed similarity of the old distinctive face to the 16 old faces without distinctive features.

Table 2.

Summed Similarities for the HS Model.

Study	Normal			Test		
	Old (same)	Old (Changed)	New	Old (same)	Old (Changed)	New
Normal	15s	16s	16s	16Ms	15Ms	16Ms
Distinctive (same type)	16Ms	3Ms	16Ms	3Bs	4Bs	4Bs
Distinctive (other type)		12Ms		12Ds	12Ds	12Ds
Target exemplar	1	M		C	M	
Total	15s+16Ms+1	16s+15Ms+M	16s+16Ms	16Ms+3Bs+12Ds+C	15Ms+4Bs+12Ds+M	16Ms+4Bs+12Ds

We estimated the six free parameters (C , D , M , B , k , and s) for each experiment by minimization of the error sum of squares to average probability of judging a face to be old (see Table 3). Note that for all three experiments the parameter D was equal to 1, indicating no reduction in similarity due to mismatching distinctive features. If D is less than 1, then the model underpredicts the number of old responses to new faces with distinctive features and overpredicts the number of old responses to new faces with distinctive features.

Table 3.

Best-Fitting Values of the HS Model's Parameters for 1-, 5-, and 10-Second Exposure Duration.

Parameters	Exposure Time		
	1 sec	5 sec	10 sec
M	0.833	0.806	0.525
s	0.015	0.008	0.006
D	1	1	1
k	1.160	0.714	0.476
C	2.043	1.510	1.450
B	2.043	1.510	1.450

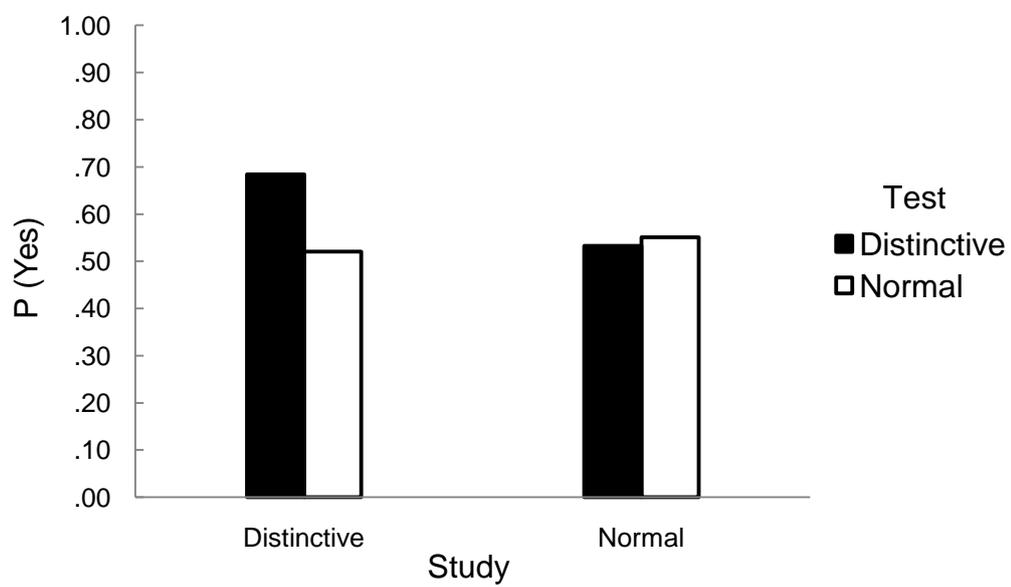
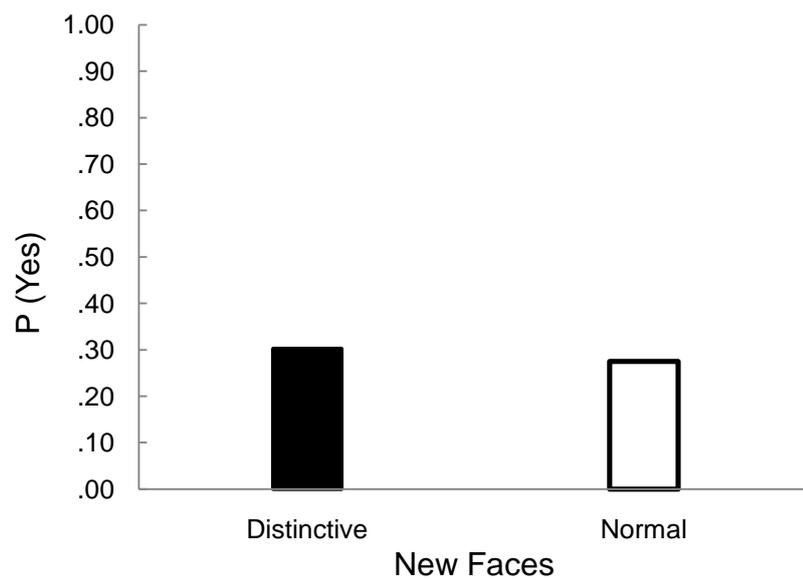
Note: M = reduction in similarity due to missing distinctive feature; B = boost in similarity due to shared similar distinctive features; C = boost in similarity due to shared identical distinctive features; D = reduction in similarity due to mismatching distinctive features; s = mean similarity of the faces; k = response criterion.

Interestingly, the parameter M slightly decreases as exposure time increases. This means that, as exposure time increases, there is a bigger reduction in similarity between a face without a distinctive feature and a face with a distinctive feature. It seems that participants, having a few extra seconds to process the faces, they become

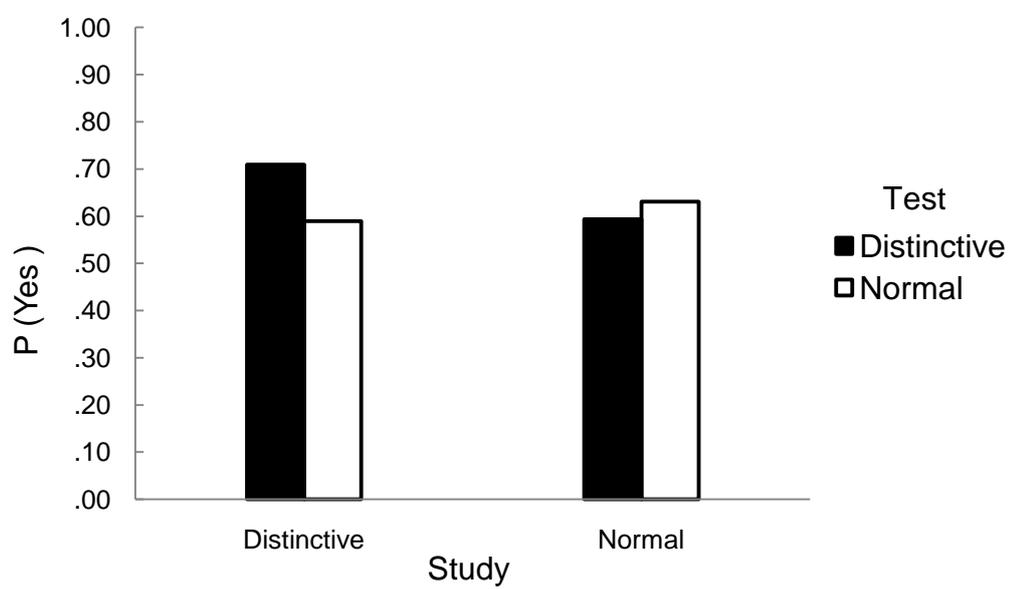
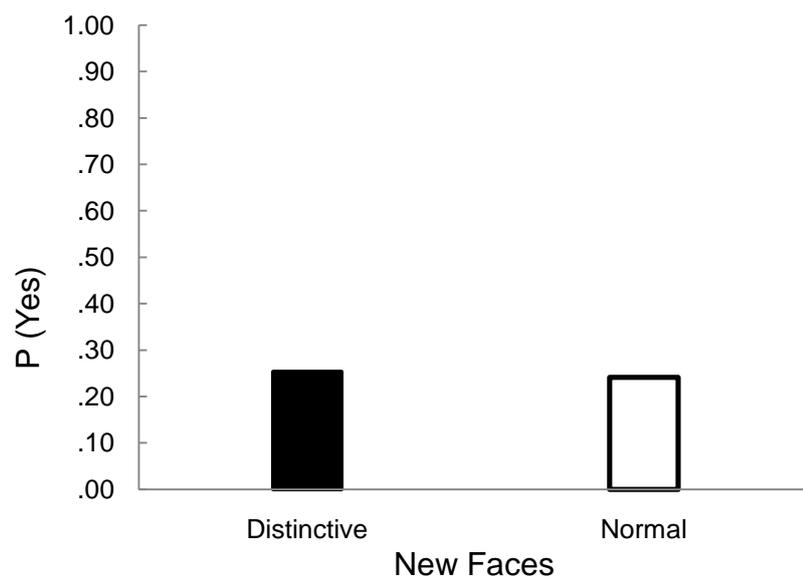
more aware of the missing distinctive feature. The C parameter, on the other hand, slightly decreases with exposure time, indicating that the boost in similarity due to matching distinctive features is higher for shorter exposures. This makes intuitive sense. As mentioned in the introduction, according to the novel popout effect, participants' attention is drawn towards unusual stimuli (i.e., the distinctive feature), which are used as highly predictive components for the recognition task that follows. Seeing another face whose distinctive feature matches the distinctive feature of another face boosts the similarity between these two faces dramatically. Based on the observed values of parameters M and C we might assume then that, when participants are only briefly exposed to faces, they are better at detecting common features, whereas as exposure time increases, they are then able to detect what is different between the faces. This is also reflected on the s parameter, which decreases with exposure time indicating that faces seem more similar to one another when participants have less time to process them. Finally, the parameter B proved to be identical to parameter C indicating no difference in how participants perceived similar and identical distinctive features. Figure 8 illustrates the predictions of the HS model for (a) 1-, (b) 5-, and (c) 10-second exposure duration. The model was successful at predicting the qualitative pattern of the results of all three experiments.

However, we might predict that in a study where the average similarity of faces would be high, the effect of removing a distinctive feature might lead to increases in old judgments because, although the mismatch with the actual study face would be reduced, the effect of not having a distinctive feature would increase similarity to the other, normal study faces. So, maybe in a study where the proportion of normal faces at study would be high, removing a distinctive feature could be predicted to create more old judgements compared to retaining a distinctive feature.

(a)



(b)



(c)

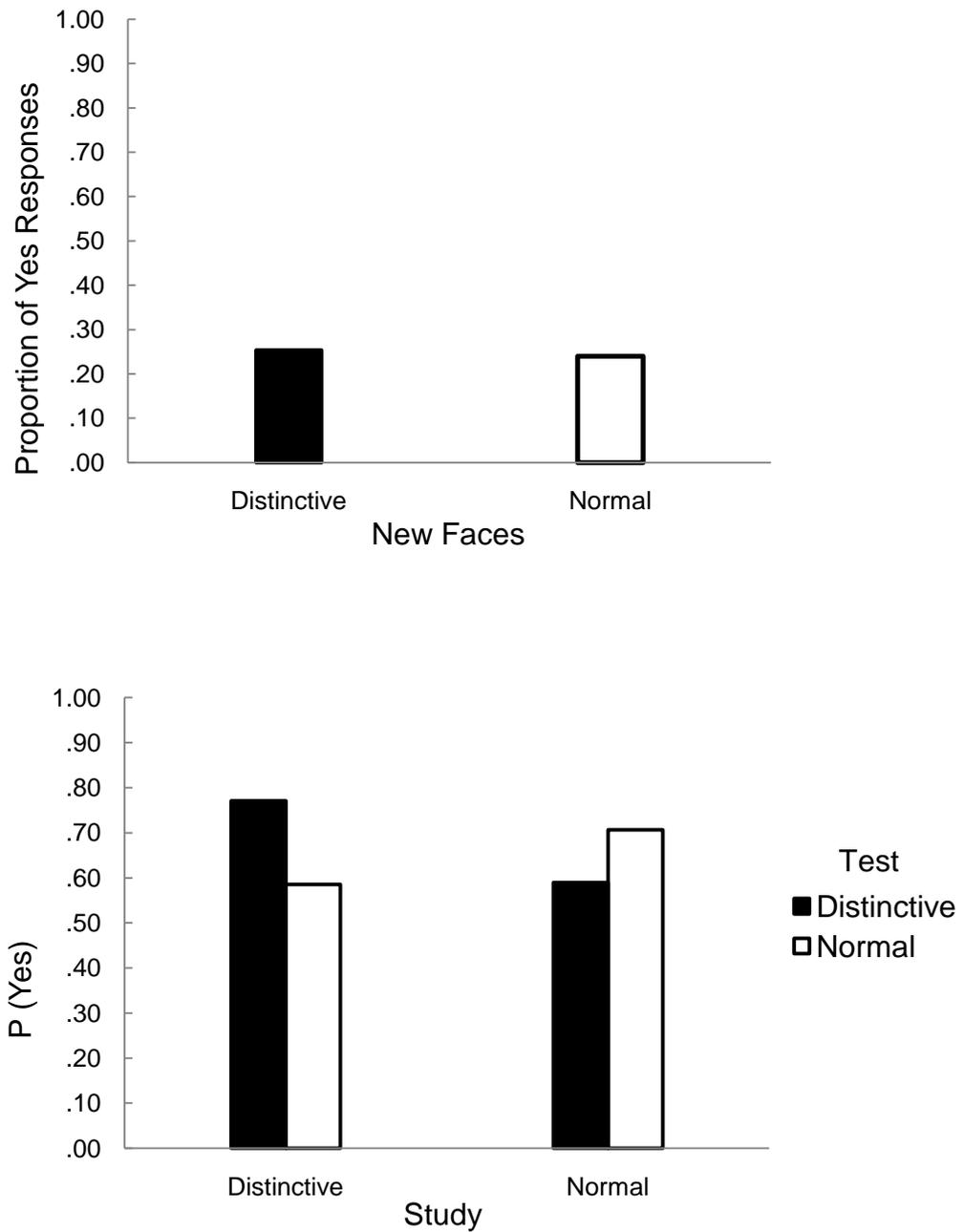


Figure 8. The HS model's predictions for (a) 1-second, (b) 2-second, and (c) 3-second exposure duration.

Discussion

The purpose of this research was to reveal the best technique to be applied when a police suspect has a distinctive feature. We hypothesized that faces that retain their distinctive feature at test are more likely to be recognised than faces that lose their distinctive feature at test. Three crossover-recognition-memory experiments confirmed this hypothesis. This difference between faces that remain distinctive and faces that lose a distinctive feature was significant independently of the exposure time manipulations. The practical implications of this result are clear. When police officers deal with a suspect with a distinctive feature and have to decide which special lineup procedure they should apply (replication or concealment), replication should be the preferred option.

A second important finding indicates that replication could be applied even for the cases where the culprit was not initially seen with a distinctive feature or the eyewitness had failed to see it during the criminal act or report it in his/her description. All three experiments revealed no difference between faces that remained without distinctive features and faces that gained a distinctive feature during test. This means that regardless of the initial encoding conditions, if a police suspect has a distinctive feature, replication will increase the likelihood of a positive identification (if the distinctive feature was present during the criminal act) compared to concealment, or will not harm identification performance (if the distinctive feature was not present during the criminal act) compared to concealment.

The use of replication will also not harm identification performance in culprit-absent lineups; across two experiments (5 seconds, and 10 seconds) new faces with distinctive features and new faces without distinctive features were equally likely to be rejected as old. Only in the 1-second-exposure experiment, new

distinctive faces were more likely to be falsely recognised as previously seen but we suspect that this result was due to the high frequency with which distinctive faces were presented to participants during the study phase. Fifty percent of the faces presented during the study phase were distinctive with distinctive features similar to the ones of the new faces. Therefore, a distinctive new face resembled in memory many other distinctive faces previously seen. The fact that new faces with distinctive features were significantly more likely to be judged as previously seen under the 1-second exposure possibly reflects the novel popout effect hypothesis: Participants, having only 1 second to view a face, their attention was immediately drawn towards the distinctive feature. Since the feature was similar to previous ones, participants judged the face as previously seen because they did not have the time to process other facial characteristics detrimental to identification.

The HS model was able to account for the qualitative pattern of our results. Faces that remained distinctive were predicted to be more likely to be recognized than faces that lost their distinctive feature. This result does not make intuitive sense, because a face that lost a distinctive feature at test should resemble in memory many faces without distinctive features seen at study, hence the probability of this face to be judged as previously seen should be high. A face with a distinctive feature at test should resemble only few faces seen at study, hence the probability of this face to be judged as previously seen should be low. However, the presence of the retained distinctive feature at test in faces that remain distinctive increases the self-similarity of the face to an extent that the overall similarity exceeds the similarity of the faces without distinctive features seen at study. This is the core assumption of the HS model.

A Comparison with Deffenbacher et al. A study by Deffenbacher et al (2000), using the same crossover-recognition paradigm, revealed a different pattern of results expressed by a higher false-alarm rate for normal instead of distinctive faces and a significant drop in recognition performance for faces that gained a distinctive feature at test compared to faces that lost a distinctive feature at test. However, it is important to note that Deffenbacher et al. used the construct of distinctiveness with a completely different meaning than we did. Their distinctive faces were caricatures generated by manipulating the similarity of a typical face relative to an average face. In our study though, distinctiveness refers to the presence of distinctive features, not to the distance of the distinctive faces from the foils in the multidimensional space.

This crucial difference may be the reason for this difference in the results. Faces in Deffenbacher's study are more likely to be perceived holistically. The presence of distinctive features on the faces in our study though, might have changed the level of processing from holistic to elemental, making the recognition task harder when the distinctive feature is absent at test. This might have happened unconsciously because of the popout effect discussed earlier, or voluntarily, because of the specific instructions that were given to participants. They were asked to examine each face carefully because they would be asked questions about them in the second part of the experiment. Therefore, a distinctive feature (e.g., a scar) on a face might have served as a highly predictive component used by the encoding system to facilitate recognition of this face at a later stage. However, at test, in the absence of this distinctive feature, the recognition system is forced to rely on more generic face information, like the shape of the face; a component that was not well encoded because of the switch of attention to the distinctive feature. In this situation, the face is judged for its overall familiarity and since there is absence of matching

distinctive features with any of the previously seen faces the probability of judging this face to be old decreases dramatically.

Locus of effect. One of the aims of Study 1 was to examine whether the distinctiveness effect operates during the encoding process, during the retrieval process, or during both processes. Our results revealed a significant interaction between study and test. Being presented with a face that remained the same between study and test produced a significantly higher hit rate compared to the cases where the face changed from one format to the other. This finding is in line with the encoding specificity phenomenon and with studies on the permanence of the distinctive features, showing that non changed faces are more accurately recognized than faces that have undergone changes in facial features between encoding and recognition phases.

Signal detection analysis. The signal detection analysis that was performed, aimed at discriminating between stimulus-based effects and criterion-based effects. The value of d' refers to the recognisability of the faces used in the particular experiment. The analysis revealed a significant interaction between study and test, which shows that participants found it significantly more difficult to discriminate faces that changed format between study and test compared to the cases where faces remained in the same format between study and test.

The analysis of the C response criterion though, showed that participants tended to be significantly more conservative (respond No more often) when faces changed format and especially when faces changed from distinctive to normal. There was no difference in response criterion between faces that remained without distinctive features and faces that gained a distinctive feature at test.

The goal of this research was to explain how the current different methods of constructing photo lineups for suspects with distinctive features lead to specific types of identification errors. These experiments were a first step toward that goal but it is clear that many details of the relationship between distinctive features and identification performance are still unknown. Further research should aim at investigating whether the current results can be replicated in a lineup-identification paradigm.

Study 2

Creating Fair Lineups for Suspects with Distinctive Features

Abstract

In their descriptions, eyewitnesses often refer to a culprit's distinctive facial features. However, in a police lineup, selecting the only member with the described distinctive feature is unfair to the suspect and provides the police with little further information. For fair and informative lineups, the distinctive feature should be either replicated across foils or concealed on the target. In the present experiments, replication produced more correct identifications in target-present lineups—without increasing the incorrect identification of foils in target-absent lineups—than did concealment. This pattern, and only this pattern, is predicted by the hybrid-similarity model of recognition.

Introduction

Imagine that you witness a crime and the culprit has an obvious marking on his forehead. You would probably feel confident that you could easily identify the culprit from a lineup at a later time. Imagine now that, using your description, the police arrest an innocent man with a similar marking on his forehead. They present you with a photo lineup in which only one person has a marking similar to the one you hold in your memory. Would you identify the innocent suspect as the perpetrator?

Eyewitness research shows that when an innocent suspect matches an eyewitness's description, errant identifications are more likely to occur when the foils do not match the description than when the foils do match the description. Put another way, an innocent suspect who stands out in a lineup is likely to be falsely identified as the culprit (Wells et al., 1998). In simultaneous lineups, in which the individuals are presented all together, eyewitnesses tend to use a relative judgment strategy (Wells, 1984; Wells et al., 1998). In this strategy, the person most closely matching the suspect is selected, even if the overall match is not good. Thus, an innocent suspect with a distinctive feature in common with the culprit is more likely to be selected when he or she is the only person in the lineup with that feature. Even if the suspect is actually the culprit, selecting the suspect from a lineup in which only he or she has the distinctive feature reported by the eyewitness, offers the police little in the way of new information. After all, the police already know about the distinctive feature from the eyewitness's description, and the eyewitness may be selecting the suspect on the basis of this old information alone.

Identification tests usually consist of a photo array or a video lineup, and police officers typically use one of two techniques to ensure that these lineups are

fair and informative. One technique is to replicate the suspect's distinctive feature across lineup members (*replication*), and the other is to conceal the area of the distinctive feature on the face of every lineup member, including the suspect (*concealment*). Both techniques ensure that the suspect does not stand out because of his or her distinctive feature. Although police officers use these procedures daily, and 34% of lineups in England and Wales are digitally manipulated in these ways because the suspects have distinctive features (P. Burton, West Yorkshire Police, personal communication, November 3, 2008), to our knowledge there is no empirical research on the effects of either technique on identification accuracy. Currently, there is no standard regulation giving preference to one technique over the other in the United Kingdom or United States. Rather, the police officer responsible for each case decides how to construct the lineup that will be presented to eyewitnesses. In Wogalter, Malpass, and McQuiston's (2004) survey of 220 jurisdictions in the United States, 77% of police officers reported replicating distinctive marks across foils, 23% reported adding similar marks to the foils, and 18% said they had tried to conceal the area of the markings. Surprisingly, 30% answered that they did nothing about distinctive features in some cases.

Both replication and concealment make the identification task more difficult for eyewitnesses, as they must rely solely on their memory of other specific facial features. But which technique allows the police to extract more information from an eyewitness's memory and therefore improve identification performance?

Nosofsky and Zaki's (2003) hybrid-similarity (HS) model of recognition predicts better performance under replication than under concealment. The HS model is a general model of the effects of distinctive features on recognition memory and has been applied to face recognition (Knapp, Nosofsky, & Busey, 2006); thus, it is

well suited to modeling these effects. When participants are asked to decide whether they have seen a particular face before, they assess the face's *familiarity*, and this judgment of familiarity determines the probability with which the participants will decide that they have in fact seen the face before. In the HS model, familiarity is defined as the total similarity between the test face and each of the exemplar faces in memory. Similarity between two faces is a joint function of their distance in a large multidimensional space (after Nosofsky, 1986) and their number of shared and unshared discrete features (after Tversky, 1977). Thus, two faces will be similar if they are near one another in the face space, have many discrete features (e.g., scars) in common, and have few unshared discrete features.

Under replication and under concealment, the target face will be, on average, more similar to the exemplars than will a foil. This is because the target matches the exemplar formed when the target was first encountered (hereafter, the *target exemplar*). Therefore, for both techniques, familiarity of the target is higher than familiarity of the foils. However, replication of features across foils at the test exaggerates this difference in familiarity between the target and the foils. Specifically, in the HS model, the common distinctive feature provides a multiplicative boost in the similarity between the target and the target exemplar and also provides a multiplicative boost in the similarity between the foils and the target exemplar. Thus, the absolute difference between the similarities of the target and the foils is increased. Conversely, concealing the target's distinctive features at the test attenuates the difference in familiarity between the target and the foils. So, when the target and foil familiarities are summed with the general familiarity to other faces in memory, the ratio of target familiarity to foil familiarity is higher for replication than for concealment. In summary, replication should increase the difference in

familiarity between the target and the foils, whereas concealment should reduce this difference. The HS model, therefore, predicts better performance under replication than under concealment in target-present (TP) lineups. Because common features only boost similarity and missing features only attenuate similarity, the HS model cannot predict the opposite pattern.

In two experiments, we compared replication with concealment. During a study phase, participants viewed a series of faces, a small proportion of which had a distinctive feature. During the test phase, a series of six-person lineups was presented. Experiment 1 used only TP lineups, and participants were forced to select a face. Experiment 2 included target-absent (TA) lineups, and participants were allowed to make a no-identification decision.

Stimuli

The stimuli were developed specifically for this study using photographs of 140 inmates from Florida's Department of Corrections Web site. The selected inmates were 24 years old and had short, brown hair and brown eyes. They were wearing the Department of Corrections uniform and were looking directly toward the camera, exhibiting neutral expressions. The photos showed only inmates' head and neck and were taken against a uniform gray background. None of the inmates wore glasses, and we removed all facial hair, bruises, scars, blemishes, moles, or other identifiers using Adobe Photoshop CS2. We then randomly selected 60 faces and digitally added a distinctive feature to each face using Photoshop. Figure 1 illustrates the six types of distinctive features that we used (a bruise, a tattoo, a piercing, facial hair, a scar, or a mole).

Prior to the experiments, 30 independent judges rated the distinctiveness, attractiveness, and degree of emotional arousal elicited by the 200 faces (80 faces in

non distinctive form only, plus 60 faces in both distinctive and non distinctive forms). We measured distinctiveness and attractiveness on 9-point Likert scales from 1 (*not at all*) to 9 (*very*). To measure emotional arousal, we used the Self-Assessment Manikin Scale (Bradley & Lang, 1994).

Of the 80 faces that never appeared with distinctive features, we excluded 4 that were outliers on the distinctiveness scale. Of the 60 faces used in both forms, we excluded 6 that were outliers on the distinctiveness scale. There were no outliers on the other scales. We also excluded 2 faces for which there was no difference in distinctiveness before versus after the addition of the distinctive feature.



Figure 1. Examples of faces used in Experiments 1 and 2 before (top) and after (bottom) the digital addition of a distinctive feature (from left to right: a bruise, a mole, a piercing, a moustache, a scar, and a tattoo).

Experiment 1

Method

Participants. We recruited 110 students (mean age = 25.5 years, $SD = 6.3$; 45% female) from the University of Warwick, and they participated voluntarily or received £2 (a little more than U.S. \$3). In a within-participants design, participants were presented with both replication and concealment lineups.

Procedure. In the study phase, participants were informed that they would view 32 faces drawn randomly from the stimulus set and would subsequently be tested on their memory of these faces. Participants were asked to view each face carefully. Of the 32 study faces, 6 randomly selected faces had distinctive features (one of each type) and appeared as targets in the test phase. The remaining 26 faces shown during the study phase appeared without distinctive features and were not seen again. The 32 study faces were presented in random order. Each face stimulus was displayed in the center of a computer screen for 2 s.

In the test phase, which followed a 5-min anagram-solving filler task, participants completed a lineup-identification task. They viewed a series of six 6-person lineups and were required to indicate which 1 member of each lineup they had seen in the study phase, indicating their choice by clicking on that member's photo with the computer mouse; they did not have the option of not responding. Participants were instructed that a person previously seen might have a different appearance at test and that their task was to recognize the person previously seen, not the exact photograph. Three of the lineups applied replication (see Figure 2a), and three applied concealment (see Figure 2b). The five fillers for each lineup were new, previously unseen faces randomly drawn from the stimulus set. Lineups were displayed in two rows of three photos each (see Figure 2). The placement of the target in each lineup was determined randomly for each participant, and the six lineups were presented in a random order, which was also determined separately for each participant. There was no time limit for making a decision, and no feedback was provided. The duration of the experiment was approximately 10 min.

(a)



(b)



Figure 2. Examples of (a) a replication lineup and (b) a concealment lineup presented in Experiments 1 and 2.

Results

Figure 3 shows the proportion of correct and incorrect selections in the two conditions. Participants were significantly more likely to correctly select the suspect

when distinctive features were replicated across foils rather than concealed on the target, $t(109) = 5.32, p < .001, p_{rep} = .99, r = .45$.

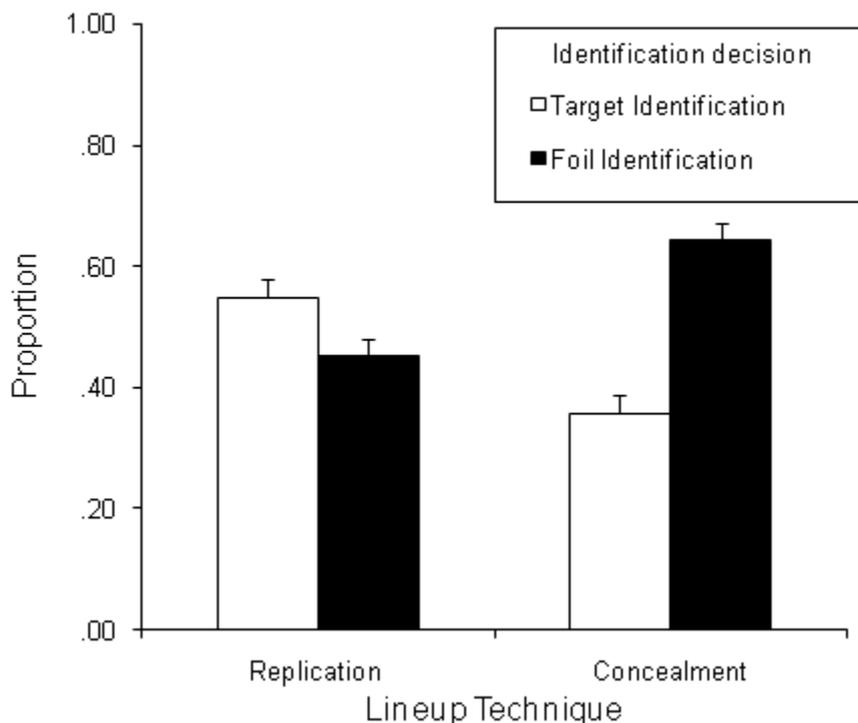


Figure 3. Mean proportion of correct responses and errors for replication and concealment lineups in Experiment 1. Error bars represent standard errors of the mean.

Experiment 2

Experiment 2 replicated Experiment 1 and extended the design to include TA lineups. The design was a 2 (lineup technique: replication, concealment) \times 2 (target presence: present, absent) within-participants design. In the TA lineups, all six foils were on average, equally familiar under replication and under concealment (because none of them matched any of the exemplars exactly), so the HS model predicted no difference in identification accuracy between the two conditions for TA lineups.

Method

Participants. A total of 85 psychology students (mean age = 20 years, $SD = 3.0$; 74% female) from the University of Warwick participated for course credit.

Procedure. The procedure was identical to that of Experiment 1, with two modifications. First, in the test phase, participants viewed 12 lineups instead of 6; half were TP and half were TA lineups. Second, if participants recognized none of the faces in the lineup, they were instructed to click on a “none” button below the lineup. TP and TA lineups were randomly intermixed.

Results

Figure 4 shows the proportion of correct and incorrect responses for the replication and concealment techniques. In TP lineups, participants were more accurate at identifying the suspect when distinctive features were replicated across foils rather than concealed, $t(84) = 5.02, p < .001, p_{rep} = .99, \underline{r} = .48$; this result replicates the results of Experiment 1. Also, the proportion of errors that were foil identifications (as opposed to no identifications) was higher in the replication condition than in the concealment condition, $t(84) = 2.74, p < .01, p_{rep} = .97, \underline{r} = .29$. In TA lineups, accuracy did not differ between the concealment and replication conditions: Similar proportions of participants incorrectly selected an innocent foil.

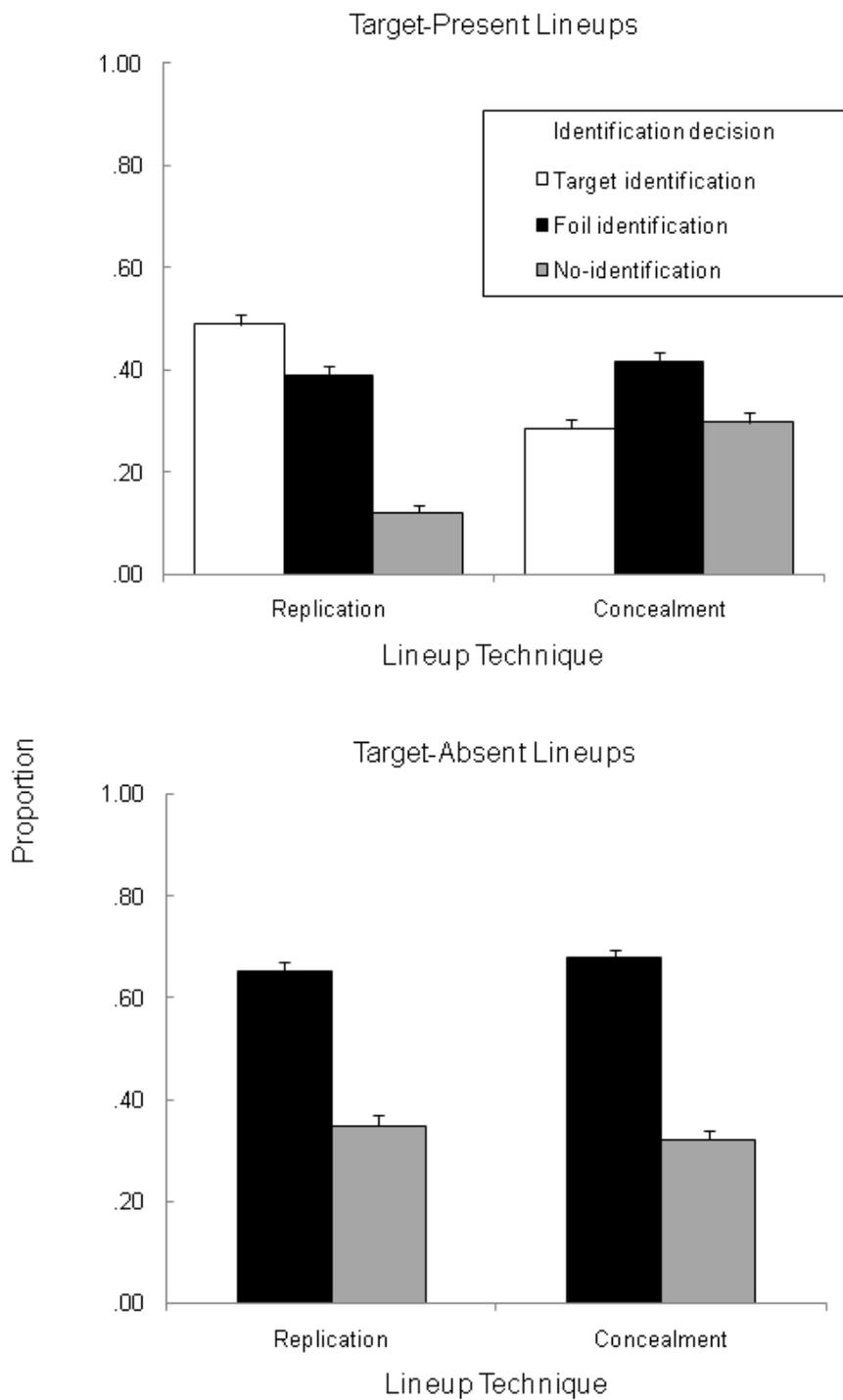


Figure 4. Mean proportion of correct responses and errors for replication and concealment lineups in Experiment 2: (a) target-present lineups and (b) target-absent lineups. Error bars represent standard errors of the mean.

In summary, our results suggest that replication is better than concealment for constructing lineups because replication increased the probability of selecting the target when a target was present without increasing the probability of selecting an innocent foil when the target was absent. The only drawback was that when the target in TP lineups was not identified, replication (compared with concealment) rendered participants less willing to make a no-identification decision. However, in absolute terms, incorrect foil selections were equally likely for the two techniques.

Discussion

Our finding that correct identifications increased in TP lineups created using replication supports the HS model of recognition memory. Standard global-familiarity models, which do not take into account the effects of distinctive features (e.g., Valentine & Ferrara, 1991), cannot account for our data. Under these models, the target:foil familiarity ratio in TP lineups is the same for concealment and replication lineups. Therefore, standard global-familiarity models predict no difference in identification performance between the two kinds of lineups. This prediction is not supported by our results.

Standard global-familiarity models also predict that participants will make increased false identifications in TA lineups created using the concealment technique: Because faces without distinctive features resemble many other faces without distinctive features seen in the study phase, the overall familiarity evoked should be increased under concealment; hence, participants shown a concealment lineup should have an increased tendency to choose someone from the lineup and to make false identifications. Under replication, the opposite should be true. However, our data revealed no difference in choice rates between replication and concealment lineups that did not include the target.

In Experiment 2, the improvement when distinctive features were replicated rather than concealed came from a reduction in incorrect no-identification decisions. It could be argued that the increase in hits in the replication condition resulted from an increased tendency to select someone from the lineup. Such a mechanism, though, would also generate more false identifications in both TP and TA lineups in the replication condition. However, in both the TP and TA lineups of Experiment 2, participants were as likely to select a person from the lineup in the replication condition as in the concealment condition, despite the fact that targets were correctly identified more often in the replication condition.

Our finding that replication (in which case the suspect remains unchanged between study and test) produces more accurate identifications than concealment (in which case the suspect is altered between study and test) is consistent with the changed-appearance literature. Lineup-identification studies, for instance, show that disguises (Cutler, Penrod, & Martens, 1987a, 1987b; Cutler, Penrod, O'Rourke, & Martens, 1986), changes in hair style or facial hair, and the addition or removal of glasses (Read, 1995) impair identification performance. Likewise, recognition-memory studies show that disguises, changes in pose or facial expression, addition or removal of glasses (Patterson & Baddeley, 1977) changes in visual angle (Bruce, 1982), and the effect of the target's aging (Read, Vokey, & Hammersley, 1990) increase false-identification rates (see also Shapiro & Penrod, 1986).

Our study is directly relevant to cases in which an eyewitness reports a culprit's distinctive feature. Wells and his colleagues argued that when a suspect has a distinctive feature that is not reported, lineups should follow the principle of "propitious heterogeneity" (Luus & Wells, 1991; Wells et al., 1998); that is, the distinctive feature should not be replicated among the foils. However, research

suggests that replication should still be applied in such cases. People are able to encode information without concurrent awareness of what is being encoded (Shanks & St. John, 1994). So, although eyewitnesses may not verbalize the presence of a distinctive feature, they may be able to remember it should they see it on the culprit at the time of the lineup. For reasons of fairness, then, everyone in the lineup should have the distinctive feature.

We used a mathematical model of the effect of distinctive features on recognition memory to make predictions for real-world lineups. We predicted that replicating a distinctive feature across foils is better than concealing it on the suspect, because replication amplifies the difference in the familiarity of the target and the foils, whereas concealment attenuates this difference. Two experiments confirmed this prediction. Police officers should be aware of this theoretical and empirical result when constructing lineups for suspects with distinctive features and should replicate rather than conceal these features.

Study 3

Lineup Construction for Suspects with Distinctive Features: To Replicate, Remove, or Pixelate?

Abstract

Replicating a suspect's distinctive feature across lineup members produces more correct identifications in target-present lineups than removing the area of the distinctive feature on the suspect's face. The present study tested another technique currently used by the police to prevent suspects from standing out: the pixelation of the area of the distinctive feature across lineup members. Experiment 1 compared replication, removal, and pixelation in a laboratory-based recognition experiment. Participants viewed faces of which a small proportion had distinctive features, and then had to identify old faces in a series of lineups. Replication produced more correct identifications than did removal or pixelation, without increasing the false-alarm rate. Experiment 2 compared replication and pixelation using an eyewitness-identification paradigm: participants were faced with a single lineup after they viewed a culprit with a distinctive feature in a staged live event. Again, replication was better than pixelation. The hybrid-similarity model of recognition predicted these results.

Introduction

Eyewitness evidence is cited as the number one cause of miscarriages of justice (Scheck, Neufeld, & Dwyer, 2001). In the US, faulty eyewitness evidence played a role in over 70% of wrongful convictions in DNA exoneration cases (www.innocenceproject.org). In England and Wales, a staggering one in five of all eyewitnesses makes a known mistaken identification (Valentine, Pickering, & Darling, 2003). It is unsurprising that eyewitness researchers have been extensively investigating the circumstances under which people are wrongfully convicted. A major finding in the literature is that suspects who stand out in a lineup are more likely to be wrongfully selected as the culprit (Wells et al., 1998). In many cases though—approximately 34% in England and Wales (Peter Burton, West Yorkshire Police, personal communication, November 3, 2008)—a suspect has some sort of distinctive facial feature, such as an unusual hairstyle, a tattoo, a scar, or a mole and the police have to decide how best to create a fair lineup.

Wells et al. (1998, p. 630) argue that “the suspect should not stand out in the lineup or photospread as being different from the distractors based on the eyewitness’s previous description of the culprit or based on other factors that would draw extra attention to the suspect” and this rule has been included in the US Department of Justice’s guide of best practice for handling eyewitness identification evidence (Technical Working Group on Eyewitness Evidence, 1999). According to Wogalter, Malpass, and McQuiston’s (2004) survey of 220 jurisdictions in the United States, 77% of police officers prevent suspects from standing out in lineups by replicating their exact distinctive features to every foil in the lineup, 23% add similar distinctive features to the foils, and 18% conceal the area of the markings. Thirty percent of respondents said that they do not do anything about distinctive

features. What we do not know is how these techniques affect identification performance. In this paper we ask how different techniques for preventing suspects from standing out in lineups affect the likelihood that witnesses select the target, or a foil, from a lineup.

In the UK, the code of practice (Code D; Home Office, 2008) enforced by amendments to the 1984 Police and Criminal Evidence Act (hereafter, *PACE*) states that when a suspect has some sort of distinctive facial feature such as a tattoo, facial hair, or an unusual hairstyle, which does not appear on the other lineup members, the distinctive feature should be replicated across lineup members (*replication technique*). If replication is not practical, then the distinctive feature should be concealed on the suspect by either removing it, pixelating it, or blocking it out with a solid black rectangle (*concealment technique*). If the distinctive feature is pixelated or blacked out, then the corresponding area on all of the lineup members is also pixelated or blacked out. Replication and concealment serve to ensure that the suspect does not stand out in the lineup. It is at the identification officer's discretion to choose which technique to apply, but *PACE* requires the decision and the rationale behind it to be recorded. When a culprit's distinctive feature is not reported by the eyewitness but the suspect has a feature deemed by the police to be distinctive, *PACE* states that concealment should be applied (*PACE Codes of Practice, Code D, Annex A, paragraphs 2A, 2B, 2C*).

The methods used to apply replication and concealment, as well as the rationale behind using one technique over another, varies across UK police forces. For instance, the Northumbria Identification Unit conducted 2,496 lineups in 2008, 40% of which involved suspects with distinctive features and in each of these cases concealment was applied as a less costly and time-consuming method (Karl Burns,

personal communication, January 30, 2009). In the same year, Hampshire Police conducted 2,800 lineups, 33% of which involved either concealment or replication (Karen Miller, personal communication, March 29, 2009). Perhaps, most importantly, these statistics show that a substantial minority of lineups require alteration to deal with suspects with distinctive features.

At the time of writing this paper, we are aware of only one study that has systematically examined the techniques that police use to prevent suspects from standing out in lineups (Zarkadi, Wade, & Stewart, *in press*). Zarkadi et al. used Adobe Photoshop to digitally alter face-photographs to include a distinctive feature. At study, participants viewed a series of faces, a small proportion of which had a unique distinctive feature. After a short filler task, participants viewed a series of lineups and identified, for each lineup, which one face (if any at all) was previously seen. In half of the lineups the distinctive feature was replicated across all foils and in the other half the distinctive feature was removed from the suspect's face. Replication increased the probability of selecting the target when the target was in the lineup without increasing the probability of selecting an innocent foil when the target was not in the lineup.

Comparing replication and feature-removal was a theoretically interesting manipulation—it enabled Zarkadi et al. (*in press*) to test predictions made by the hybrid similarity (HS) model of recognition memory (Knapp, Nosofsky, & Busey, 2006; Nosofsky & Zaki, 2003). But the police often use a different concealment technique to hide distinctive features: as mentioned above, they pixelate the area of the distinctive feature on the suspect's face and the corresponding area on the other lineup members. For this reason, in Experiment 1, we adapted Zarkadi et al.'s design and used it to compare three techniques: replication, removal, and pixelation. As in

Zarkadi et al., we used the HS model to make predictions about participants' performance.

According to the HS model, to judge whether a test face has been seen before, people compare the test face to the faces they have in memory from the study phase (hereafter, the *exemplars*). The probability that a previously-seen test face will be accurately identified is determined by the sum of the similarity between the test face and each of the exemplars; the higher the summed similarity, the higher the familiarity feeling for this particular test face, hence the higher the probability of an accurate identification.

The degree of similarity between two faces depends on their distance in a large multidimensional space (after Nosofsky, 1986) and on the counts of the number of shared and unshared, discrete features (after Tversky, 1997). So two faces will be similar if they are close together in the multidimensional space, have many features in common, and few unshared, discrete features. The two forms of similarity are combined multiplicatively; that is, to be similar, faces must be both close in the multidimensional space and have matching feature sets.

Under all three techniques (replication, removal, pixelation), the target face is, on average, more similar to the exemplars than is a foil because the target matches the exemplar formed in memory when the target was encountered in the study phase (hereafter, the *target exemplar*). Therefore, familiarity of the target is higher than familiarity of the foils, regardless of which technique is used. Under replication, however, the difference in familiarity between the target and the foils is exaggerated. The common distinctive feature boosts multiplicatively the similarity between the target and the target exemplar and also boosts multiplicatively the similarity between the foils and the target exemplar. Thus the absolute difference between the

similarities of the target and the foils is increased by replication. Conversely, under removal, the difference in familiarity between the target and the foils is attenuated because the common distinctive feature has been eliminated. So, when these familiarities are combined with the general familiarity to other, background faces, the target:foil familiarity ratio should be higher for replication, and lower for removal. This pattern should be true of all target-present (TP) lineups. The HS model, therefore, clearly predicts better performance under replication than under removal. However, the HS model does not make clear-cut predictions about which form of concealment—pixelation or removal—will produce more correct identifications. Predictions depend upon the relative sizes of the multiplicative boost from the match between a pixelated feature and the original feature and the match between a removed feature and the original feature. In target-absent (TA) lineups, performance should be equal under all three techniques because all six foils are, on average, equally familiar—none of them matches exactly any of the exemplars.

Experiment 1

Method

Participants. Ninety-five University students (mean age = 28.9 years, $SD = 8.7$, 53% female) participated voluntarily. All participants were tested in all six conditions of a 3 (lineup technique: replication, removal, pixelation) x 2 (target-presence: present, absent) within-participants design.

Stimuli. The stimuli were developed especially for this experiment using the faces of 148 inmates from Florida's Department-of-Corrections website. All selected inmates were 24-26 years old, with short brown hair and brown eyes, and wore the Department of Corrections' uniform. The photos were in colour, taken against a uniform blue background, and showed the inmates' head and upper torso as they

looked directly at the camera. None of the inmates wore glasses, and we removed all tattoos, scars, moles, and other identifiers using Adobe Photoshop CS3.

Of the 148 faces, 31 faces were left unaltered, and 117 randomly selected faces were digitally altered using Photoshop to each include one distinctive feature (either a scar, a bruise, an eyebrow cut, a blemish, a mole, some facial hair, crossed eyes, a birthmark, or a tattoo). Thus, there were 13 faces for each of the nine types of distinctive feature. Next we pixelated the area of the distinctive feature to create a third, pixelated version of each face. Thus, the final set consisted of 117 faces without distinctive features, the same 117 faces with distinctive features, the same 117 faces with the area of the distinctive feature pixelated, and 31 additional faces without a distinctive feature that were used as fillers in the study phase. Figure 1 provides examples of faces from all three versions.



Figure 1. Example face used in Experiment 1 before (from left to right), after the digital addition of a tattoo, and after the pixelation of the area of the tattoo.

Procedure. In the study phase, participants were told that they would see a series of 40 faces, one at a time for 5 s each, and subsequently their memory for these faces would be tested. Each trial started with a fixation cross presented for 300 ms centred on the screen. Next a face stimulus appeared centred on the screen.

Participants were asked to inspect each face carefully. Of the 40 faces, nine had

distinctive features (one of each type) and were drawn randomly from the set of faces with distinctive features. These nine faces appeared in the subsequent test phase as targets. The remaining 31 faces had no distinctive features and were not seen again. The 40 faces were presented in random order.

In the test phase, which followed a 5-minute anagram-solving filler task, participants completed a lineup-identification task. Participants were told that they would view a series of 18 six-person lineups and their task would be to indicate for each lineup which one lineup member (if any) was shown in the study phase by clicking on it with the mouse. If participants believed that none of the lineup members had been shown before, they were instructed to click on a “none” button below the lineup. They were instructed that a person previously seen might have a different appearance at test and that their task was to recognize the person previously seen, not the exact photograph. Six of the lineups were created using the replication technique (Figure 2a), six were created using the removal technique (Figure 2b), and six were created using the pixelation technique (Figure 2c). For each type of lineup, half contained the target (TP lineups) and half did not (TA lineups). The fillers for each lineup were new, unseen faces. Lineups were displayed in two rows of three photos (Figure 2). The placement of the target in the lineups was random and the 18 lineups were presented in a different random order for each participant. There was no time limit for their decision and no feedback was provided. The duration of the experiment was approximately 15 minutes.

(a)



(b)



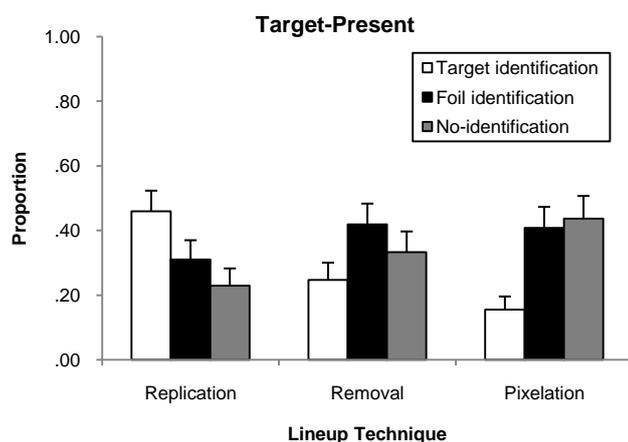
(c)



Figure 2. Examples of a (a) replication-technique lineup, (b) a removal-technique lineup, and (c) a pixelation-technique lineup presented in Experiment 1.

Results

(a)



(b)

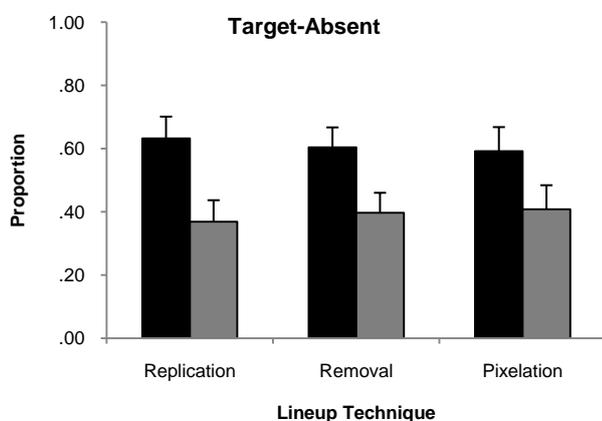


Figure 3. Mean proportions of correct responses and errors under replication, removal, and pixelation for (a) TP lineups (a) and (b) TA lineups. Error bars represent the standard error of the mean.

Figure 3 shows the proportions of correct and incorrect responses for replication, removal, and pixelation in TP lineups (Figure 3a) and TA lineups (Figure 3b). A 3 (lineup technique: replication, removal, pixelation) x 2 (target presence: present, absent) ANOVA on correct responses revealed a main effect of lineup technique, $F(2, 188) = 16.48, p < .001$. There was also a main effect of target-

presence, $F(1, 94) = 10.82, p = .001$, and a significant interaction between lineup technique and target-presence, $F(2, 188) = 23.54, p < .001$. To consider separately the differences among lineup techniques for TP and TA lineups, a simple main effects model was examined separately for TP and TA lineups.

TP lineups. There were significant differences among lineup techniques for TP lineups, $F(2, 188) = 42.55, p < .001$. To evaluate the three pairwise differences among the means for TP lineups, planned contrasts were performed using t-tests with a Bonferroni-adjusted alpha of .017. Replication elicited significantly higher target-identification rates than either removal, $t(94) = 5.79, p < .001$ (two-tailed), $r = .51$, or pixelation, $t(94) = 8.81, p < .001$ (two-tailed), $r = .67$. Finally, removal elicited significantly higher target-identification rates than pixelation for TP lineups $t(94) = 2.91, p = .005$ (two-tailed), $r = .29$. Looking at participants' errors, the proportion of errors that were foil identifications (rather than no-identifications) was similar under replication and under removal, $t(94) = .01, p = .99$, but higher under replication compared to pixelation, $t(94) = 2.65, p = .03$ (two-tailed), $r = .26$, and higher under removal compared to pixelation, $t(94) = 2.72, p = .03$ (two-tailed), $r = .27$.

TA lineups. Finally, a one-way repeated-measures ANOVA on correct responses revealed no significant differences among lineup techniques on TA lineups, $F(2, 188) = .18, p = .84$; across the three lineup techniques, participants were equally likely to correctly respond no one or to choose an innocent foil.

We conclude that replication is better for constructing lineups than both removal and pixelation because replication increased the probability of selecting the target when the target was present, and it also decreased the probability of selecting an innocent foil or making a no-identification decision. In line with our prediction, this advantage for replication occurred without increasing foil identifications in TA

lineups: across the three lineup techniques participants were almost equally likely to select an innocent foil. These results are in line with the Zarkadi et al.'s study (in press) and the predictions of the HS model.

As this pattern of results has been demonstrated only in laboratory-based face recognition studies, we next sought to examine whether these results could be replicated in a more ecologically valid, eyewitness-identification paradigm.

Experiment 2

In Experiment 2 we used a different lineup identification paradigm, in which participants viewed a single target (a confederate) with a scar in a staged event, and after 24 hours they were asked to identify the culprit from a lineup. To maximize experimental power, in Experiment 2 we compared only replication (the optimal technique in Experiment 1) and pixelation (the technique of focus in this paper). Replication and pixelation were tested in both TP and TA lineups. We predicted that the results from Experiment 1 and Zarkadi et al.'s study (in press) would generalize to this eyewitness-identification paradigm.

Another aim of Experiment 2 was to obtain confidence ratings from participants before they viewed the lineup (*pre-lineup confidence*) and after they viewed the lineup (*post-lineup confidence*). Research has shown that eyewitnesses' statements of confidence about defendants' guilt have a huge impact on juries (Bradfield & Wells, 2000). If, as research has shown (Vokey & Read, 1992), people believe that distinctive faces are easier to remember than non-distinctive faces, then pre-lineup confidence should be relatively high. However, the identification task will be very hard for eyewitnesses because under both replication and pixelation, they will have to rely solely on specific facial features—other than the scar that they hold in memory—to make an accurate identification. For this reason, we also predicted a

drop in confidence after viewing both types of lineups and we expected post-lineup confidence to be a better predictor of identification accuracy than pre-lineup confidence. Furthermore, if, as the HS model posits, the target:foil familiarity ratio is higher under replication than under pixelation, then post-lineup confidence should be higher for correct responses under replication than under pixelation in TP lineups. Therefore, the drop between pre- and post-lineup confidence should be smaller under replication. Following the same reasoning, confidence ratings for incorrect responses should be lower under replication than under pixelation. Overall, confidence ratings should be a better predictor of identification performance under replication than under pixelation in TP lineups. If, as the HS model posits, in TA lineups all six foils evoke equivalent levels of familiarity for both techniques, there should not be a difference in post-lineup confidence between replication and pixelation.

Method

Participants. We recruited 123 people (mean age = 33.9 years, $SD = 8.0$, 37% female) from offices on the University campus to participate voluntarily. Participants were departmental administrators, academics, researchers, or doctoral students.

Design. The design was a 2 (lineup technique: replication, pixelation) x 2 (target presence: present, absent) between-participants design. Participants were randomly assigned to one of the four lineup conditions.

Materials. Thirty-six lineups were created especially for this study. We recruited a confederate who would take part in the staged event, and thus would be the target of the subsequent lineups (Figure 4). To create the foils for the lineups, we asked 5 students to describe the confederate shortly after he asked them for road directions. On the basis of these descriptions, we selected 11 students (all white, 18-21 year-old male undergraduates) who agreed to be photographed as foils.

Photographs of these students (and the confederate) were taken in front of a neutral background, with all students looking straight to the camera, in neutral expression, and all wearing the same sweater. Like the confederate, all had short, light brown hair, a medium build, and they were clean-shaven. After conducting a mock-witness test (Doob & Kirshenbaum, 1973) with 20 independent judges, we excluded 2 foils that were selected as the target at an above-chance level. This left 9 foils.



Figure 4. Photo of the confederate in Experiment 2.

For each of the 10 photographs (1 confederate and 9 foils), three versions were created: one normal, one with a scar on the left cheek under the eye, and one digitally altered to pixelate the area of the scar. The scars were fake tattoo transfers, and produced a near identical effect when applied to the face of each lineup member. None of the 20 judges who participated in the mock-witness test found any photo unusual due to the transfer tattoo; they only found great coincidence in the fact that every lineup member had a similar scar on their left cheek.

Each lineup consisted of nine close-up colour photos. The reason we decided to include 9 lineup members (instead of six) was to accord with the current guidelines of PACE, which state that the lineup should include at least eight other people apart from the suspect (PACE, Codes of Practice, Code D, Annex A, Para. 2). Each lineup applied either the replication technique (Figure 5a) or the pixelation

technique (Figure 5b). Both TP and TA lineups were used. The order of the photographs in each lineup was random.

(a)



(b)



Figure 5. The two lineup techniques tested in Experiment 2: (a) replication, and (b) pixelation.

Study phase. To recruit participants, we followed Levi's (2007) methodology. The experimenter and the confederate (with scar) visited people in their offices. The confederate introduced himself and the experimenter and asked whether they would be willing to take part in a two-minute psychological experiment that would take place in their office on the following day. If the person agreed, the confederate asked whether they would be available on the following day at the same time. Only in the

cases where the participant wanted a different time, a different, mutually convenient time was arranged and the confederate wrote down the time for the experiment. After that the participant was thanked and the experimenter and confederate left the office.

Each time slot lasted for approximately 5 minutes. For this reason the offices of successive participants were near each other. The next appointment was typically arranged for 24 hours later (approximate range = 24-26 hours).

Test phase. For each test-phase, the experimenter visited participants' offices in the same order (where possible) that she had visited them during recruitment. This time she was without the confederate, and she asked each participant to identify from a lineup the person who accompanied her the previous day. She stated clearly that this person might or might not be in the lineup. Before showing them the lineup, she asked them how likely they thought, on a scale from 0% to 100%, that they would make a correct decision.

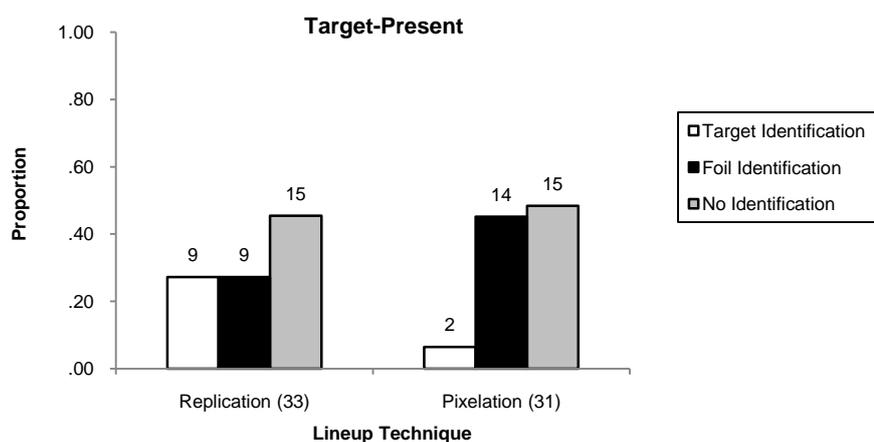
Next, the experimenter showed the participant the lineup, and participants either selected a person or made a no-identification judgment. Participants were asked to give a confidence rating about their decision, on a scale from 0% to 100%, then they were debriefed and thanked for their participation.

Results

Identification performance. Figure 6 shows the proportions of correct and incorrect responses for each lineup technique (replication, pixelation) in TP lineups (Figure 6a) and TA lineups (Figure 6b). To examine the effect of each variable on accuracy, we performed a hierarchical log-linear analysis with lineup technique (replication, pixelation) and target-presence (TP, TA) as factors. In TP lineups a correct response is target identification. In TA lineups a correct response is a no-identification. In TA lineups participants could only make a mistake by identifying a

foil. However, in TP lineups, participants could make a mistake either by selecting no one or by identifying a foil. For this reason, at this initial stage of analysis, we combined TP participants' incorrect responses, such that the log-linear analysis compared correct responses to incorrect responses.

(a)



(b)

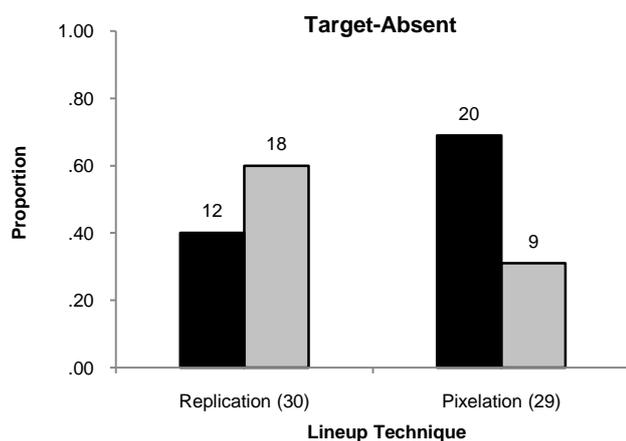


Figure 6. Proportion of participants who made target identifications, foil identifications, and no identifications as a function of lineup technique (replication vs. pixelation) in (a) TP lineups and (b) TA lineups. Data labels are absolute frequencies. N per condition is in brackets.

The final model produced by this analysis retained the lineup technique x result (i.e., correct vs. incorrect) interaction and the target-presence x result interaction. The likelihood ratio of this model was $\chi^2(2) = 1.42, p = .49$. The lineup technique x result interaction was significant, $\chi^2(1) = 8.87, p = .003$. This interaction indicates that participants' overall performance (correct versus incorrect responses) was different for replication than for pixelation. The target-presence x result interaction was also significant, $\chi^2(1) = 11.99, p < .001$. This interaction indicates that participants' overall performance was different in TP lineups and TA lineups.

Looking at the performance on TP lineups in more detail, the distribution of target-identifications, foil identifications and no identifications across the two lineup-techniques were examined using chi-square analysis. A two-way chi-square analysis comparing the frequency of target identifications and no identifications did not indicate a significant effect of lineup-technique. However, a two-way chi-square analysis comparing the frequency of target-identifications and foil-identifications indicated a significant effect of lineup-technique, $\chi^2(1) = 5.44, p = .02, \phi = .40$, with more target identifications under replication than pixelation, and more foil identifications under pixelation than under replication. Based on the odds ratio participants were 7 times more likely to select an innocent foil under pixelation than under replication. A two-way analysis of data from the TA condition indicated that the frequency of foil identifications was higher under pixelation than under replication, $\chi^2(1) = 4.98, p = .03, \phi = .29$, which is in line with the log-linear analysis reported earlier. Based on the odds ratio, participants were 3.33 times more likely to select an innocent foil under pixelation than under replication.

Confidence-Accuracy Relationship. Another objective of Experiment 2 was to investigate whether there are significant differences in the relationship between

eyewitness confidence and accuracy under replication and under pixelation.

Participants' percent distribution of self-reported pre-lineup confidence is illustrated in Figure 7. Foil identifications and no identifications are collapsed and labelled "incorrect". Pre-lineup ratings were significantly higher for correct as compared with incorrect responses, $t(121) = 3.26$, $p < .001$, $r = .31$, showing that pre-lineup confidence was a good predictor of accuracy.

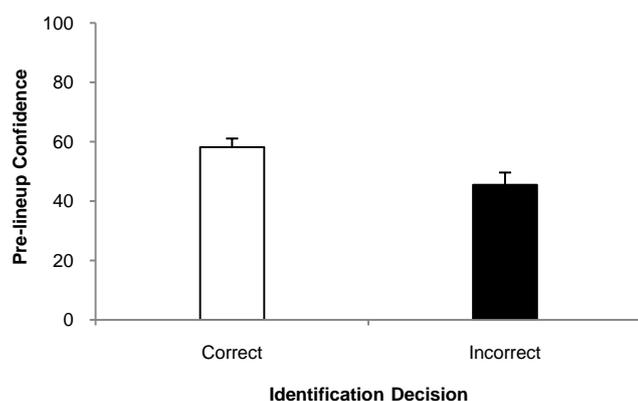


Figure 7. Participants' percent distribution of self-reported pre-lineup confidence.

Error bars represent the standard error of the mean.

Post-lineup confidence ratings were analyzed separately for TP (Figure 8a) and TA (Figure 8b) lineups given the differences in post-lineup confidence-accuracy relationship that have been reported in the literature (e.g., Malpass & Devine, 1981). For this purpose, a series of univariate ANOVAs were conducted. In TP lineups, a 2 (lineup technique: replication, pixelation) x 3 (identification decision: target identification, foil identification, no identification) ANOVA conducted on post-lineup confidence ratings revealed a main effect of lineup technique, $F(1,58) = 13.87$, $p < .001$, $\omega^2 = .22$. In line with our prediction, participants who had viewed a replication-technique lineup rated themselves significantly more confident in their

lineup decision as compared to those who had viewed a pixelation-technique lineup. There was also a main effect of identification decision, $F(2, 58) = 7.93, p < .001, \omega^2 = .23$. Participants who made a correct decision rated themselves significantly higher in confidence as compared to those who made a wrong decision. Post hoc Tukey tests revealed higher confidence for target identifications as compared with foil identifications and for target identifications as compared with no identifications. There was no difference in confidence for no identifications and foil identifications. The interaction between identification decision and lineup technique was not significant, $F(2, 58) = .06, p = .94$. In TA lineups, in line with our prediction, a 2 (lineup technique: replication, pixelation) x 2 (identification decision: foil identification, no identification) ANOVA revealed no significant differences in post-lineup confidence ratings between replication and pixelation, $F(1, 55) = .04, p = .84$. However, there was still a main effect of identification decision, $F(1, 55) = 50.06, p < .001, \omega^2 = .88$. Post-lineup confidence ratings were higher for no identifications as compared with foil identifications. No significant interaction between lineup technique and identification decision was observed, $F(1, 55) = 2.13, p = .15$.

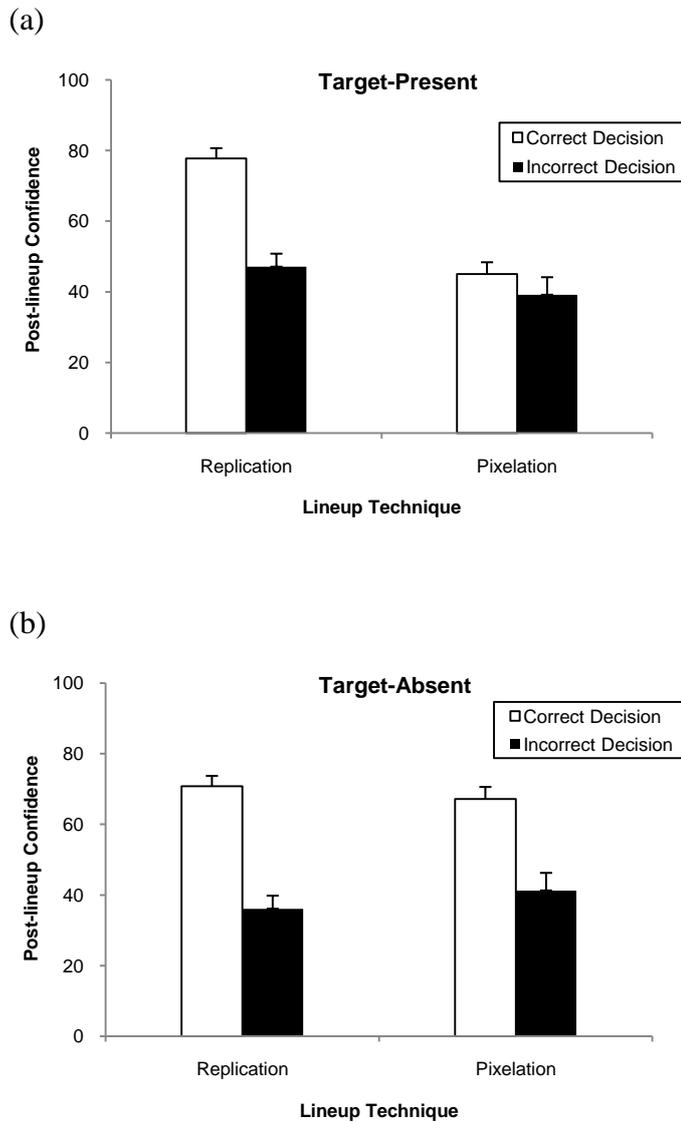


Figure 8. Participants' percent distribution of self-reported post-lineup confidence for replication and pixelation in (a) TP lineups and (b) TA lineups. Error bars represent the standard error of the mean.

Changes in confidence ratings were also examined separately for TP (Figure 9a) and TA (Figure 9b) lineups. For TP lineups, a 3 (identification decision: target identification, foil identification, no identification) x 2 (lineup technique: replication, pixelation) ANOVA conducted on the difference in confidence before versus after viewing the lineup revealed a marginally significant main effect of lineup technique,

$F(1, 58) = 3.99, p = .05, \omega^2 = .05$, revealing a tendency of the participants to increase their confidence rating after viewing a replication-technique lineup, but not after a pixelation-technique lineup. This result is in line with our prediction. There was also a main effect of identification decision, $F(2, 58) = 6.35, p = .003, \omega^2 = .18$. Post hoc Tukey tests revealed higher increase in confidence for target identifications as compared with foil identifications and for target identifications as compared with no identifications. There was no difference in confidence increase for no identifications and foil identifications. The interaction between identification decision and lineup technique was not significant, $F(2, 58) = 2.30, p = .11$. For TA lineups, a 2 (identification decision: no identification, foil identification) x 2 (lineup technique: replication, pixelation) ANOVA conducted on the change in confidence before versus after viewing the TA lineup revealed no main effect of lineup technique, $F(1, 55) = 1.51, p = .22$. In line with our prediction, the change in confidence was similar for participants who had viewed a replication-technique lineup and for those who had viewed a pixelation-technique lineup. However, there was a main effect of identification decision, $F(1, 55) = 9.77, p = .003, \omega^2 = .16$. Participants' confidence ratings were significantly increased after viewing the lineup when their lineup decision was correct as compared to when their lineup decision was incorrect. The interaction between lineup technique and identification decision was not significant, $F(1, 55) = 1.63, p = .21$.

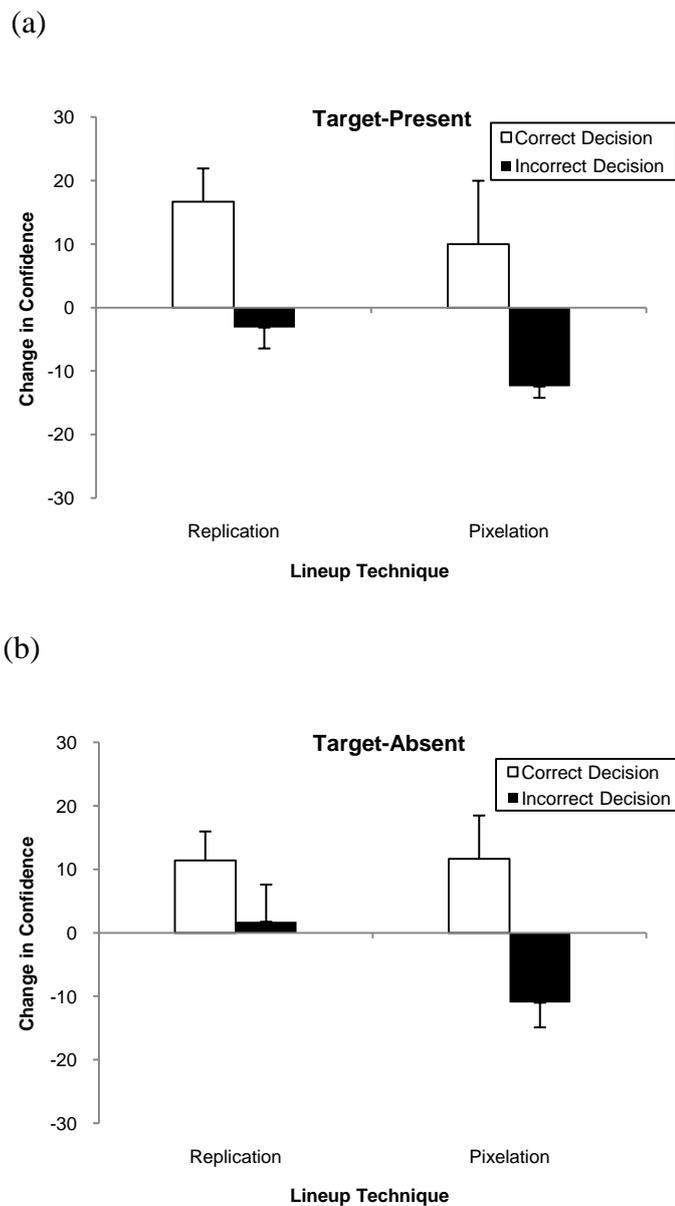


Figure 9. Participant's percent distribution of change in self-reported confidence before versus after viewing (a) a TP lineup and (b) a TA lineup. Error bars represent the standard error of the mean.

We conclude that replication proved to be a better technique for constructing lineups than pixelation, because replication increased the probability of selecting the target when the target was present, and consequently decreased the probability of

selecting an innocent foil or making a no-identification decision. However, contrary to expectation, in TP lineups, although participants under replication were more confident in their decision than were those under pixelation, the ability of post-lineup confidence to predict identification performance did not differ between replication and pixelation participants.

An accuracy advantage for replication was also apparent in TA lineups: foil identifications were less likely under replication than under pixelation. In line with our prediction though, in TA lineups, there was no difference in post-lineup confidence ratings between replication and pixelation.

Discussion

The central finding of this research was that replicating the suspect's distinctive feature across lineup members is better than both pixelating the area of the suspect's distinctive feature on the foils (Experiments 1 and 2) and removing the distinctive feature from the face of the suspect (Experiment 1). Indeed, higher target-identification rates in replication lineups are not just the outcome of an increased tendency for participants to select someone from a replication lineup because choosing rates in TA lineups were equal across all three techniques. We argue that the increased target identifications are a result of the increased difference in familiarity between the target and the foils under replication.

Removal proved to be better than pixelation (Experiment 1), showing that pixelating a feature on a face should be regarded as adding a new, different distinctive feature over the top of the original feature, rather than concealment of the original feature. Within the HS framework (Nosofsky & Zaki, 2003), this result must mean that a face with a scar (for example) is more similar to an unscarred version of

the face (i.e., a face with the scar removed) than to a face with the scar pixelated.

This relative ordering of similarities seems quite intuitive with hindsight.

The results of Experiment 1 cannot be predicted by other global-familiarity models (e.g., Valentine & Ferrara, 1991). Following these models, in TP lineups the target:foil familiarity ratio is the same for all three techniques because the target is more familiar than the foils only because it matches the target exemplar. Therefore, other global-familiarity models predict no difference in identification performance under replication, removal, and pixelation. Furthermore, following other global familiarity models, in TA lineups, global familiarity should be higher under removal; hence choosing rates (i.e., foil identifications) should be higher. This is not supported by our results.

Our results are, however, in line with the literature on changed appearance. Shapiro and Penrod's (1986) meta-analysis showed that when the target's appearance changes from study to test, there is a lower probability that participants will select someone from the lineup, and if participants nevertheless choose someone, there is a higher probability of a mistaken identification. Several other studies conducted after this meta-analysis have revealed similar results (e.g., Cutler, Penrod, & Martens, 1987a, 1987b; Read, 1995; Read, Vokey, & Hammersley, 1990).

In Experiment 2, the improvement when distinctive features were replicated versus removed or pixelated was observed as an increased probability of identifying the target when he was in the lineup and a decreased probability of selecting an innocent foil. This observation replicates Zarkadi et al.'s (in press) findings. However, in Experiment 2, the advantage for replication was seen as a reduced probability of selecting an innocent foil when the target was absent, which was not apparent in Experiment 1 or Zarkadi et al. We speculate on possible causes of this

unexpected result immediately below, but note here that the result does not change the overall conclusion: replication should be preferred to pixelation or removal.

So, replication not only increases correct identification in TP lineups but, in some circumstances, can reduce incorrect foil identifications in TA lineups.

Experiment 2 differs from earlier experiments in that the target in the lineup had been experienced in real life the day before rather than as a picture on a computer screen. It is possible that, under these circumstances, participants made more abstract inferences about the person (e.g., is he honest?) instead of attempting to remember just physical appearance. Another reason for the observed reduced foil identifications could be the use of larger lineups in Experiment 2 (nine-person lineups instead of six-person lineups). Indeed research has shown that in larger lineups, there is a lower probability of the witness selecting an innocent foil in target-absent lineups (Levi & Lindsay, 2001).

A further reason for the difference is that choosing rates were lower in Experiment 2 compared to previous experiments, with lower target-identification rates and higher no-identification rates in both TA and TP lineups. This finding could reflect the increased level of task-difficulty in Experiment 2. For instance, the increased retention interval between study and test (24 hours instead of 5 minutes in previous experiments) might have increased the level of difficulty in two ways. At a cognitive level, the longer the time interval, the higher the possibility of memory decay and the more the opportunities for interference with the target's memory trace (Brewer, Weber, & Semmler, 2005). At a meta-cognitive level, it is possible that the long time interval made participants believe that they would not be able to make a positive identification. Such beliefs could lead people to raise their criteria, requiring more evidence in order to choose someone.

Another objective of Experiment 2 was to investigate whether there are significant differences in the relationship between eyewitness confidence and accuracy under replication and under pixelation. Although for many years eyewitness researchers have reported a weak relationship between accuracy and confidence in the literature (Brewer, Weber, & Semmler, 2005), in this study both pre-lineup confidence and post-lineup confidence proved to be good indicators of identification accuracy. In line with our prediction and the HS model, participants under replication rated themselves higher in confidence about their lineup decision compared to the participants under pixelation. Finally, participants raised their confidence rating significantly after they made a correct decision but there were no differences between the two techniques.

Although this paper has focused on the cases where an eyewitness refers to the culprit's distinctive feature, there are several cases in which the eyewitness does not report the suspect's distinctive feature. There are several reasons for such a situation to occur: (a) The eyewitness had not seen the distinctive feature (e.g., she only saw one profile of the culprit), (b) the eyewitness did not verbalize/recall the presence of the distinctive feature although she had consciously encoded it, (c) the eyewitness did not verbalize/recall the presence of the distinctive feature because she encoded it but without concurrent awareness of what was being encoded (see Shanks & St. John, 1994), (d) the culprit did not have the distinctive feature at the time of the criminal act, or (e) the police caught the wrong person; the culprit did not have a distinctive feature but the innocent suspect has.

For reasons (b) and (c) only, replicating the distinctive feature of the suspect across lineup members should be beneficial for the eyewitness's memory. For the rest of the reasons though, concealment might be a better option and the most

effective way to create a match-to-description lineup where the suspect does not stand out. Under these circumstances, the present results suggest that removal, rather than pixelation, is the more effective concealment technique. Nevertheless, because the police officer responsible for a case won't probably know the reasons why the eyewitness failed to report the suspect's distinctive feature, the parameters of the individual case should be carefully considered, and common sense should determine when distinctive features should be concealed or replicated. Further research should investigate this issue.

Finally, in the current study we used simultaneous lineups, that is, all six lineup-faces were presented simultaneously to participants. However, police officers often present lineups sequentially to eyewitnesses, that is, one lineup member at a time. We predict that our results would generalize to sequential lineups as long as eyewitnesses are informed that the lineup members have replicated distinctive features. However, it is not clear how eyewitnesses would use this information. This is something we are currently investigating.

Study 4

Testing Police Lineups for Suspects with Distinctive Features Using a Videotaped Simulated Crime

Abstract

It is common practice amongst US and UK police officers to either replicate or conceal a suspect's distinctive feature in an effort to protect them from standing out in a lineup. Two previous studies have revealed a superiority of the replication technique in both increasing target identifications in target-present lineups and in some cases decreasing innocent foils' identifications in target-absent lineups. This study, using a videotaped crime event, revealed a different pattern of results: concealment (with the form of removal) reduced the likelihood of choosing an innocent foil when the target was absent from the lineup. No differences in performance among the three techniques were observed in lineups where the target was present. The hybrid-similarity model of recognition cannot account for these results.

Introduction

As mentioned in the previous studies, in the US and UK, in an effort to protect suspects with distinctive facial features from standing out in a lineup, police officers either digitally replicate the distinctive feature on every lineup member (replication) or conceal it in some way on the face of the suspect (concealment). Although there are no standard regulations for using one technique over the other, concealment is usually used as a cheaper and more straightforward option. However, previous research has shown that replication is a superior technique (Zarkadi, Stewart, & Wade, 2009; Zarkadi, Wade, & Stewart, in press). This study aimed at providing further support for replication's superiority using a simulated crime but, to foreshadow the results, instead found an advantage for concealment.

Four experiments in two recent studies by Zarkadi et al. (in press) and Zarkadi et al. (2009) have systematically assessed the effect of the replication and concealment techniques on identification performance. In three face-recognition experiments, the authors used Photoshop, currently used by the police, to alter digitally the faces of current inmates from Florida's Department-of-Corrections website in order to test replication and concealment. At study, participants viewed a series of faces of which a small proportion had a unique distinctive feature. At test, participants were presented with a series of lineups and had to identify, for each lineup, which one face (if any) was previously seen. Replication was applied by replicating the target's distinctive feature across foils. Concealment was applied either as (a) *removal*, where the distinctive feature of the target was removed from his face, therefore no one in the lineup had a distinctive feature, and as (b) *pixelation*, where the distinctive feature of the target was pixelated on his face and the pixelation was replicated onto every other foil's face. Replication proved to be the best

technique, followed by removal, followed by pixelation for increasing target identifications in target-present (TP) lineups without increasing foil identifications in target-absent (TA) lineups.

In a fourth experiment (Zarkadi et al., 2009), the authors pitted replication and pixelation using an eyewitness identification paradigm: the experimenter and a confederate with scar visited offices in a University campus asking members of staff and administrators to take part in a psychological experiment that would take part in their office on the following day. After 24 hours, the experimenter visited the offices by herself and presented each eyewitness with a lineup asking them to identify the person who was with her on the previous day. Half of the participants were tested in a replication-technique lineup and half were tested in a pixelation-technique lineup. This time, not only did replication increase target-identification rates in target-present lineups, but it also reduced foil-identification rates in target-absent lineups.

The superiority of the replication technique (expressed with higher target-identification rates in target-present lineups) is explained by the HS model (Nosofsky & Zaki, 2003). The HS model predicts better identification performance under replication than under concealment. Key point in the HS model is the concept of familiarity of a test face, which is measured by the summed similarity between the test face and the old faces. So the more similar a face is, on average, to the old faces, the more familiar this face is. The degree of similarity between two faces depends on two measures: (a) their distance in a large multidimensional space and (b) the number of shared and unshared, discrete features. So the closer two faces are in the multidimensional space and the more discrete features they have in common and the fewer discrete features they have unshared, the more similar they are. So, within this framework, both under replication and under concealment, the target evokes a higher

feeling of familiarity than do the foils. Under replication though, this difference in familiarity between the target and the foils is amplified because of a multiplicative boost in similarity caused by the common distinctive feature. Under concealment however, this difference is attenuated. As a result, by combining these familiarities with the general familiarity to other, background faces, the difference in familiarity between the target and the foils is increased under replication but attenuated under concealment. So, performance under replication is predicted to be better than under concealment in target-present (TP) lineups. In target-absent (TA) lineups there is no difference in familiarity among foils under both replication and concealment, so identification performance is predicted to be equal for both conditions.

In both studies the HS model proved to be successful in modelling the qualitative pattern of results for each of the two lineup techniques in Zarkadi's et al. (in press) and Zarkadi's et al (2009) data: replication was better than removal and pixelation at increasing target-identification rates. Pixelation proved to be the worst of the three techniques suggesting that pixelated distinctive features were perceived by the experimental participants as new distinctive features that were not previously seen. Therefore, faces with pixelation were least familiar when compared to the study phases. The only result that the HS model was not able to account for was the decreased foil-identification rates under replication observed in target-absent lineups in the real-world study (Zarkadi et al., 2009).

The present experiment aimed at providing additional support to the HS model and to the results of Zarkadi et al. (in press) and Zarkadi et al. (2009) using a videotaped simulated crime instead of a face-recognition task or a live event. We expected that the same pattern of results would be revealed: overall, replication would lead to higher identification accuracy compared to both removal and

pixelation. It would also be interesting to see if the effect of replication in TA lineups would be again apparent.

Method

Participants

A total of 204 people ($M = 25.4$ years, $SD = 8.2$, 126 female, 78 male) participated either voluntarily (University students) or for £2 payment (iPoints participants).

Design

The design was a 3 (lineup technique: replication, removal, pixelation) x 2 (target presence: TP, TA) between-participants design. Participants took part online and were randomly assigned to one of the six lineup conditions. Online testing is now well-established in the area of cognitive psychology (Birnbaum, 2000).

The Video

The video simulated an event (a culprit accessing someone's computer without authorization) taking place in a University office and was filmed by the experimenter. An undergraduate psychology student (Figure 1) served as the culprit for £10 payment. The recording set the scene at a dark, empty office before showing a man entering the office, leaning in front of a computer, exposing to the camera his left profile, deleting some files from a computer, and exiting the office. The duration of the video was 20 seconds. The culprit's face was exposed for 5 seconds only while deleting the files (see Figure 1). The scar was visible for 4 seconds. Both during entering and exiting the office, the camera was recording the culprit from behind.



Figure 1. Photo of the confederate in Experiment 1.

The Lineups

The replication and pixelation lineups were taken from Zarkadi et al.'s study (2009) (see Figure 2a and 2b). To create the removal lineups, we removed digitally the scar from the replication lineups (see Figure 2c). Ten white male undergraduate students from Warwick University served as foils. The selected students were 19-20 years old, had short, light brown hair, neutral expressions, and were wearing the same green jumper. Students were looking directly towards the camera. The lineup photos showed only students' head and neck and were taken against a uniform white background. None of the students was wearing glasses and all blemishes, scars, moles or other identifiers were removed with Adobe Photoshop CS3. The student who served as the culprit in the video also fitted this description. More detailed description of the face stimuli can be found in Zarkadi et al. (2009).

(a)



(b)



(c)



Figure 2. The three lineup techniques tested in Experiment 1: (a) Replication, (b) Removal, and (c) Pixelation.

Procedure

Before the study phase, participants were informed that the experiment they would take part in was about how people interpret events. In the study phase, participants were informed that they would view a short, videotaped event that would last 20 seconds. They were asked to view the video carefully because subsequently

they would be asked some questions about it. The video was presented centered on the screen.

The test phase followed a 5-minute anagram-solving filler task. Participants were asked to indicate in a 10-point Likert scale how confident they were that they would be able to identify the person they saw in the video from a 6-person photo lineup (a score of 1 indicated *not confident at all* and a score of 10 indicated *extremely confident*) before seeing the lineup. They viewed a six-person lineup and were asked to indicate, which *one* lineup member (if any) they saw previously in the video. Participants used the mouse to select a photograph or press the “none” button; they did not have the option of not responding. Lineups were displayed in two rows of three photos each. The placement of the target in each lineup was random for each participant. There was no time limit for their decision and no feedback was provided. After participants responded, they were asked to give a confidence rating about their decision on a 10-point Likert scale (a score of 1 indicated *not confident at all* and a score of 10 indicated *extremely confident*). The duration of the Experiment was approximately 7 minutes.

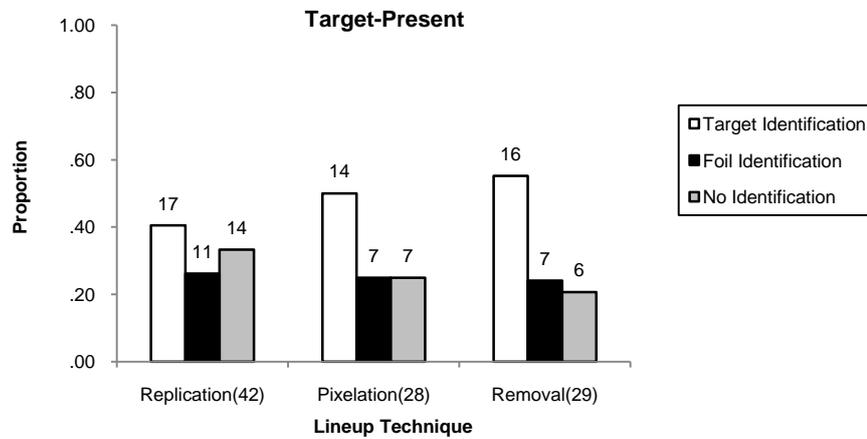
Results

Identification Performance

Figure 3 shows the proportions of hits, misses, correct rejections and false alarms for the three lineup techniques (replication, removal, pixelation) in TP and TA lineups. A hierarchical loglinear analysis was conducted with three factors: Identification decision (correct, incorrect), lineup technique (replication, concealment, pixelation), and target-presence (TP, TA). The analysis revealed no effect of target-presence. Identification performance was similar for TP and TA lineups regardless of the lineup technique. The lineup technique x identification

decision interaction was significant, $\chi^2(2) = 7.64, p = .02$, indicating that participants' overall performance (correct versus incorrect responses) was different across the three lineup techniques. The likelihood ratio of this model was $\chi^2(6) = 2.92, p = .82$ indicating that this model was a significantly good fit of the data.

(a)



(b)

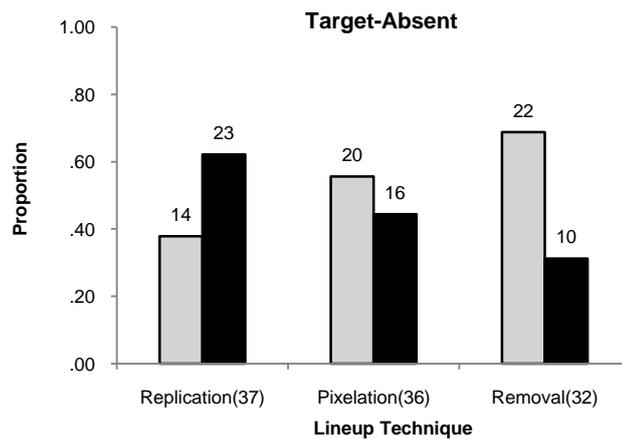


Figure 3. Proportions of correct responses and errors in replication, removal, and pixelation lineups for (a) TP and (b) TA lineups. Data labels are absolute frequencies. N per condition is in brackets.

For consistency with previous literature and, given the differences in performance between TP and TA lineups that are usually reported, we examined TP

and TA lineups separately. In TP lineups the distribution of target identifications, foil identifications and no identifications across the three lineup techniques were examined using a 3 x 3 chi-square analysis, which revealed similar number of target-, foil-, and no identifications across the three lineup techniques, $\chi^2(4) = 1.942, p = .746$. Analysis of data from the TA condition indicated that no identifications were significantly higher under removal than under replication, $\chi^2(1) = 6.571, p = .01, \phi = .309$, but not compared to pixelation, $\chi^2(1) = 1.249, p = .264$. There was no difference between the replication and the pixelation techniques, $\chi^2(1) = 2.302, p = .129$.

Confidence-Accuracy Relationship

Another objective of this experiment was to investigate whether there are significant differences in the relationship between eyewitness confidence and accuracy under replication, removal, and pixelation.

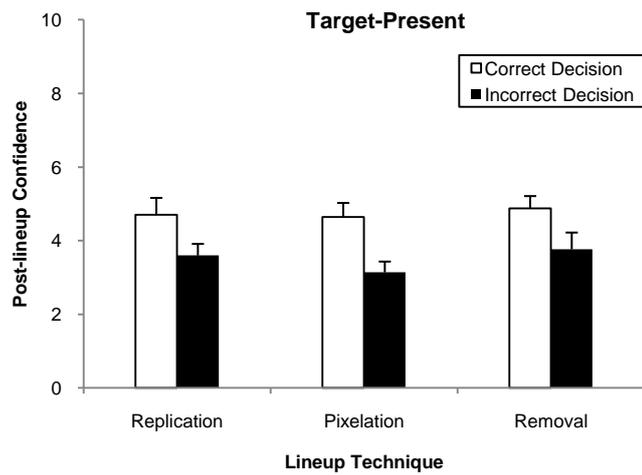
Pre-lineup Confidence. Contrary to Zarkadi et al. (2009), pre-lineup confidence ratings were not significantly higher for correct as compared with incorrect responses, $t(202) = .312, p = .755$, showing that pre-lineup confidence was not a good indicator of accuracy. At the time that participants give their confidence rating, they are not aware of the technique that has been applied to the lineup they are going to view, nor do they know whether the target is present or absent in the lineup, therefore prelineup confidence was tested only in terms of whether it could predict correct versus incorrect responses.

Post-lineup Confidence. To analyze post-lineup confidence ratings, a series of univariate ANOVAs were conducted separately for TP (Figure 4a) and TA lineups (Figure 4b). Note that for simplicity, in Figure 4a, foil identifications and no identifications are collapsed and labeled “incorrect”. In TP lineups, a 3 (lineup

technique: replication, removal, pixelation) x 3 (identification decision: target identification, foil identification, no identification) ANOVA conducted on post-lineup confidence ratings revealed a main effect of identification decision, $F(2, 90) = 7.36, p = .001, \omega^2 = .14$. In line with the Zarkadi et al. (2009) study, participants who made a correct decision rated themselves significantly more confident in comparison to those who made a wrong decision. Post hoc Tukey tests revealed higher confidence for target identifications as compared with foil identifications and for target identifications as compared with no identifications. There was no difference in confidence for no identifications and foil identifications. Contrary to Zarkadi et al., there was no main effect of lineup technique, $F(2, 90) = 0.65, p = .526$; participants under replication rated themselves equally confident with those under pixelation or under removal. As in Zarkadi et al., the interaction between identification decision and lineup technique was not significant, $F(4, 90) = .15, p = .96$.

In TA lineups, in line with our prediction, a 3 (lineup technique: replication, removal, pixelation) x 2 (identification decision: foil identification, no identification) ANOVA revealed no significant differences in post-lineup confidence ratings for replication, removal, and pixelation, $F(2, 99) = 1.44, p = .24$. Also, contrary to Zarkadi et al., there was no main effect of identification decision, $F(1, 99) = .77, p = .38$, showing that post-lineup confidence was not a good indicator of accuracy; post-lineup confidence ratings were similar for no identifications as compared with foil identifications. As in Zarkadi et al., no significant interaction between lineup technique and identification decision was observed, $F(2, 99) = .42, p = .66$.

(a)



(b)

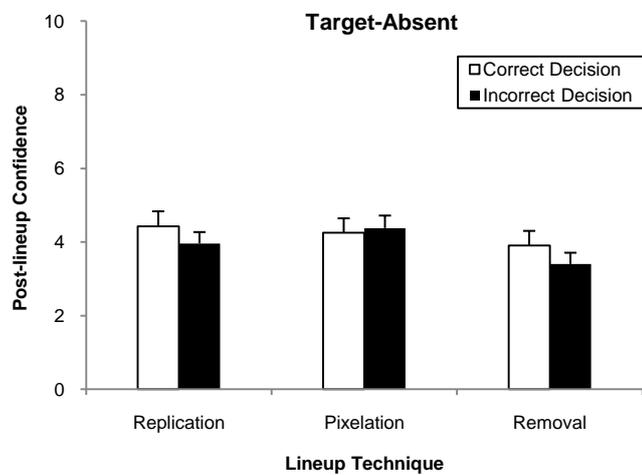


Figure 4. Participants' percent distribution of self-reported post-lineup confidence for replication, removal, and pixelation in (a) TP and (b) TA lineups. Error bars represent the standard error of the mean.

Change in Confidence. Finally, another series of univariate ANOVAs was conducted in order to analyze changes in confidence ratings separately for TP (Figure 5a) and TA (Figure 5b) lineups. For TP lineups, a 3 (identification decision:

target identification, foil identification, no identification) x 3 (lineup technique: replication, removal, pixelation) ANOVA conducted on the difference in confidence before versus after viewing the lineup revealed a main effect of identification decision, $F(2, 90) = 4.53, p = .013, \omega^2 = .08$. Post hoc Tukey tests revealed higher increase in confidence for target identifications as compared with no identifications but not for target identifications as compared with foil identifications or for no identifications as compared with foil identifications. This result is in line with Zarkadi et al. There was no significant main effect of lineup technique, $F(2, 90) = .64, p = .53$, revealing no tendency of the participants to increase their confidence rating after viewing a replication-technique lineup (Zarkadi et al. found this effect to be marginal). The interaction between identification decision and lineup technique was not significant, $F(4, 90) = .37, p = .83$.

For TA lineups, a 2 (identification decision: no identification, foil identification) x 3 (lineup technique: replication, removal, pixelation) ANOVA conducted on the change in confidence before versus after viewing the TA lineup revealed no main effect of lineup technique, $F(2, 99) = .997, p = .37$. In line with our prediction, the change in confidence was similar for participants who had viewed a replication-technique lineup, for those who had viewed a pixelation-technique lineup, and for those who had viewed a removal-technique lineup. There was also no main effect of identification decision, $F(1, 99) = .81, p = .37$. Participants' confidence ratings were equally increased after viewing the lineup when their lineup decision was either correct or incorrect. The interaction between lineup technique and identification decision was not significant, $F(2, 99) = .23, p = .796$.

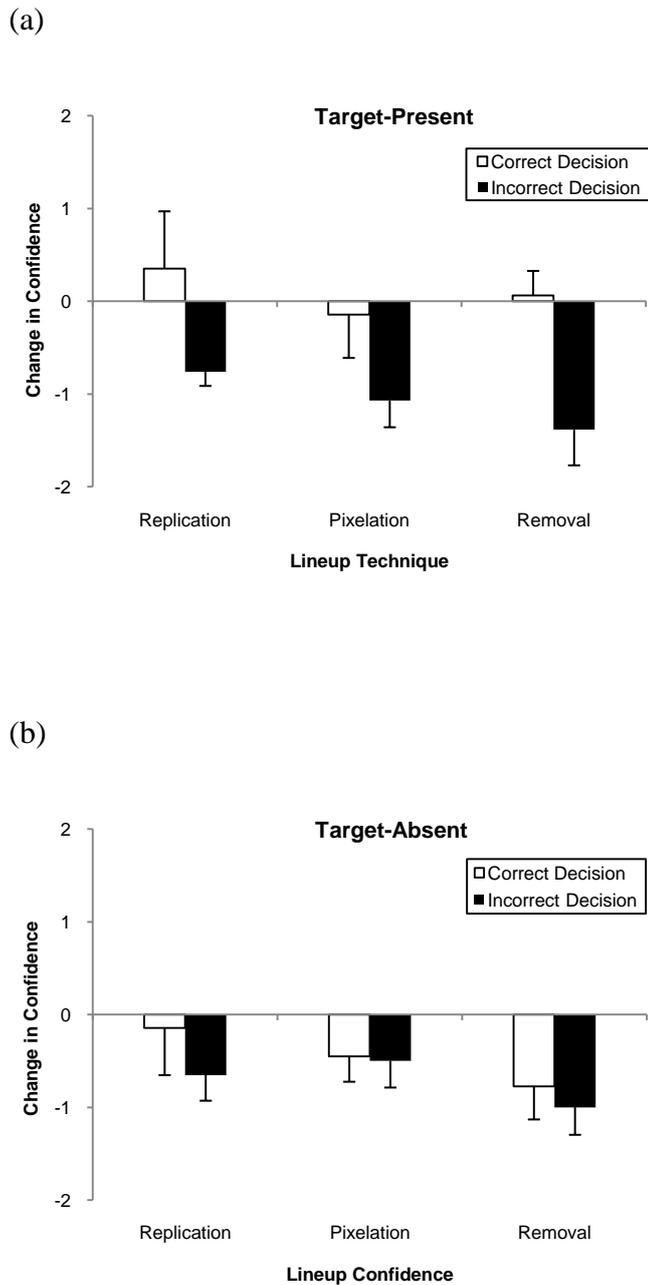


Figure 5. Participant's percent distribution of change in self-reported confidence for replication, removal, and pixelation before versus after viewing (a) a TP lineup and (b) a TA lineup. Error bars represent the standard error of the mean.

Discussion

The current experiment failed to replicate the results of previous research, where replication was shown to improve target-present identifications (Zarkadi et al.,

in press; Zarkadi et al., 2009) and to support the predictions of the HS model.

Overall, removal proved to be better technique than replication and equally effective as pixelation. Removal decreased the probability of selecting an innocent foil when the target was absent, although it did not increase the probability of selecting the target when the target was present.

Surprisingly, there was no difference in identification performance among the three techniques in target-present lineups. The HS model can account for this result, but only if it is assumed that both matching and mismatching distinctive features have no effect on similarity. That is, there is a special case of the HS model in which the boost in similarity due to matching distinctive features and the decrease in similarity due to mismatching distinctive features equals 1, which means that all faces have equal self-similarities. In fact this special case reduces the HS model to the GCM upon which it was based by ignoring the distinctive features. However, because the same face stimuli and the same distinctive features were used in the Zarkadi et al. (2009) study, there are no reasons to believe that distinctive features would have no effect. Then why did the present study failed to reveal a replication advantage in target-present lineups?

In applying the HS model, Zarkadi et al. (in press) and Zarkadi et al. (2009) have assumed that the absolute difference of the ratio of target familiarity to foil familiarity is higher for replication than for concealment. So, when the target and foil familiarities are summed with the general familiarity to the other study faces, the ratio of target familiarity to foil familiarity is higher for replication than for concealment, hence participants are less likely to select a foil under replication than under concealment. It becomes apparent that because replication and concealment give multiplicative effects, the ratio is only higher to the extent that there is a fixed

familiarity to other previously seen faces. In this experiment however, there was only one face during the study phase, which might be the cause of the null result; because there was only one study face, multiplicative boost did not affect the target:foil familiarity ratio.

Although the target:foil familiarity ratio in the case of one study face would be equal for replication and concealment, the absolute difference would still be increased under replication as a result of the multiplicative boost that the common distinctive feature causes. But to what degree can the increase of the absolute difference in familiarity explain the increased identification performance under replication in TP lineups observed in the previous studies? Although the HS model does not make any such predictions, in the context of a more deterministic rather than a probabilistic system, relative judgment processes lead to similar target-identification rates under all three techniques. The face that evokes the higher (no matter how much higher) level of familiarity will be selected more frequently than each of the faces that evoke a lower level of familiarity. So, within a more deterministic system, the HS model would predict similar rates of target identifications for all three techniques. However, this was not observed in any of the previous studies.

Another unexpected result of this experiment is the observed decrease in foil identifications in target-absent lineups under removal. Following the HS model, Zarkadi et al. (in press) and Zarkadi et al. (2009) have assumed that, since the target:foil familiarity ratio is equal for both techniques in target-absent lineups, there should be no difference in identification performance: foil-identification rates should be similar for replication and concealment. Based on this logic, the HS model fails to account for the better identification performance under removal in target-absent

lineups in the present experiment. It rather predicts similar foil-identification rates among all three techniques in target-absent lineups; under all three techniques, all foils are, on average, equally familiar. Under other global-familiarity models though, in target-absent lineups, global familiarity is higher under replication because participants have seen only one face and this face had a distinctive feature, hence choosing rates (i.e., foil identifications) should be higher under replication. Therefore, other global familiarity models can account for the better performance under removal in target-absent lineups.

As an alternative explanation, one could argue that the overall familiarity evoked by a replication lineup is higher than the overall familiarity evoked by a removal/pixelation lineup; each lineup member with the replicated distinctive feature evokes a higher level of familiarity compared to the familiarity evoked by a lineup member with the distinctive feature removed/pixelated. Therefore, one could argue that replication lineups are more likely to lead to more foil-identification decisions because participants would tend to choose someone more often under replication. However, if this was true, it should also be observed in target-present lineups too. However, such an effect was not observed; replication led to similar (not higher) choosing rates as removal and pixelation. Also the results of the face-recognition studies of Zarkadi et al. (in press) and Zarkadi et al. (2009) do not support this global-familiarity argument either: first, choosing rates were equal in target-absent lineups for all lineup techniques, which means that people exhibit the same tendency to pick someone regardless of the technique applied, and second, in target-present lineups the increase in choosing rates under replication, was a result of the higher discriminability of the target face (increased choosing rates were exhibited as higher target-identification rates). Both these observations provide support for the presence

of a relative judgment strategy applied by the participants while making an identification decision, which is in line with previous studies (Wells, 1984; Wells et al., 1998). Participants seem to compare each lineup member to each other and pick the one that is more familiar compared to the rest.

From all the above, it becomes clear that different assumptions about the decision processes involved while making an ID decision lead to different patterns of results. So, instead of arguing that the HS model failed to account for these results, we might want to interpret the present findings by explaining why participants might have applied a different choosing strategy (i.e. absolute judgment) in this study. One possible reason is the fact that the task of this experiment was easier; overall identification performance was higher in this experiment compared to Zarkadi et al. This might have been because of several reasons. For example, in this experiment there was a shorter interval between study and test, there was only one study face, and the face was viewed from a closer distance. Also, the target might have been processed more meaningfully than in previous studies because participants were informed that the experiment was about how people interpret events, and not about memory. More meaningful processing usually leads to better memory performance (Lockhart & Craik, 1990). Participants possibly compared each lineup member to their memory of the target (and not to the other foils) because they had a better memory of the target. So, in target-absent lineups, under removal, each foil looked more different than the target than under replication, therefore they avoided choosing someone, and by luck, that was the correct answer. Indeed, as shown in Study 1, removal of a distinctive feature impairs memory significantly more than retention of the distinctive feature. When the target was present though, they were making correct

decisions under all three techniques because they had a very good memory of the target (for the above reasons).

Concerning the confidence-ratings data, pre-lineup confidence was not a good indicator of accuracy. Post-lineup confidence though, was higher for correct as compared to incorrect responses in TP lineups, in line with Zarkadi et al. However, the present experiment revealed no differences in confidence based on the lineup technique that was used. Finally, post-lineup confidence was not a good indicator of accuracy in TA lineups, a result that was not apparent in Zarkadi et al.

To sum up, the failure of the present experiment to replicate previous studies might have been due to any of the reasons mentioned above, or due to inadequate statistical power, or the use of online participants. However, note that this study is ecologically less valid than Study 3, the findings of which should not be overlooked. Further research investigating how other factors mediate the relationship between lineup technique and accuracy will provide a better interpretation of the present findings.

General Discussion

As shown in the introduction of this thesis, although eyewitness researchers have extensively examined the selection of foils for police lineups, relatively less attention has been devoted to the selection of foils for police lineups where the suspect has some sort of distinctive facial feature. Hence little research exists on how distinctive features influence identification accuracy and whether the presence or absence of such features could disrupt the normal mechanisms by which we recognize unfamiliar faces.

Summary and Discussion of the Experimental Findings

Study 1 attempted to observe the effect of distinctive features in recognition memory by applying a crossover-recognition paradigm. There are some preliminary findings that were already established in the literature. Previous research had shown an advantage (expressed with a lower probability of a mistaken identification) for faces that do not change format between study and test confirming the basic principle of the encoding specificity phenomenon. This finding was apparent in both recognition-memory studies (Cutler, Penrod, & Martens, 1987a, 1987b; Cutler, Penrod, O'Rourke, & Martens, 1986; Read, 1995) and eyewitness-identification studies (Bruce, 1982; Patterson & Baddeley, 1977; Read, Vokey, & Hammersley, 1990). This line of research (see Shapiro & Penrod, 1986 for a meta-analysis) had involved disguises, hair style, addition/removal of glasses, differences in pose and expression, differences in visual angle, and targets' aging effects.

Study 1 adds to this growing body of evidence indicating that faces with distinctive features (different to those that had been previously investigated) that change format between study and test are significantly less likely to be identified than faces that remain the same. The advantage of faces that do not change format

between study and test was observed across all three recognition-memory experiments and the results were independent of the exposure-time manipulations; increased exposure to the target improved recognition performance for faces that had the same format during study and during test but did not improve recognition performance for faces that changed format between study and test. This is not to say though that there may not be some critical exposure-time periods in which this effect would no longer be apparent. With much longer exposure durations, a recognition-accuracy improvement for faces that change format between study and test might be observed because participants might use the extra time to shift their attention from the distinctive feature to other facial components, detrimental to face recognition.

The most important finding of Study 1 though is that the advantage of faces that did not change format between study and test was not symmetrical. In particular, there was a decrease in recognition performance for faces that lost a distinctive feature compared to faces that retained their distinctive feature but there was no such advantage for faces that remained without distinctive features compared to faces that gained a distinctive feature during test. These two findings have important practical implications. They suggest that when the police have a suspect with a distinctive feature, regardless of whether this feature was present or not during the criminal event, and regardless of whether the eyewitness reported it in their description, replication of the distinctive feature would either not harm recognition performance compared to concealment or it would even improve it. The lower probability with which faces that lose a distinctive feature are recognised (possibly due to the eyewitnesses' adoption of a rather conservative response criterion) should also be taken into account when assessing the reliability of an eyewitness after a showup, deck identification, or a street identification. It is also an important finding to be

aware of while creating facial composites. In most instances when creating facial composites for culprits with distinctive features, the usual approach is to add the distinctive feature by drawing it on the image using a graphics package (Stuart Gibson, personal communication, February 02, 2009). Then the composite is printed-out and is sealed in an evidence bag as the eyewitness' evidence. Failure of the eyewitness to describe the distinctive feature as fully and accurately as possible will considerably decrease the likelihood of the free perpetrator's identification by members of the public (provided that the perpetrator still has the distinctive feature).

The HS model of recognition memory was able to account for the results of all three experiments in Study 1. However, it would be a strong test for the HS model to investigate whether the pattern of results would vary as a result of changes in the proportions of distinctive and non distinctive faces seen during the study phase or whether results are independent of such manipulations and are subject to different ones (e.g., manipulations of exposure time and stimulus salience).

This retention-of-the-distinctive-feature advantage observed in the experiments of Study 1 was in line with the results of the three computer-based, lineup-identification experiments of Studies 2 and 3. The aim of these experiments was to observe whether the results of Study 1 would be generalized in a lineup-identification study. It was hypothesized that under replication, where the target's appearance did not change between study and test, participants would be significantly more likely to identify the target compared to removal and pixelation lineups. The results supported the initial hypothesis. This means that replication lineups were more sensitive than concealment and pixelation lineups. Furthermore, foil identifications under replication were equally (Study 2) or less (Study 3) likely than under concealment/pixelation.

Across all three lineup-identification experiments of Studies 2 and 3, identification performance did not differ among the different techniques when the target was absent, showing that for replication, removal, and pixelation lineups, fairness is equal, and confirming the predictions of the HS model. On average, about 60% of participants chose an innocent foil from a target-absent lineup, a result that is also perfectly in line with Levi's (1998) findings from 47 experiments, according to which 60% of participants do choose an innocent foil when the target is absent.

Finally, removal led to better identification performance than pixelation (Study 3). Modelling with the HS model suggests that a pixelated feature on a face was treated by the experimental participants as a new, different distinctive feature over the top of the original feature, rather than concealment of the original feature. However, note that participants were not informed that pixelation might or might not cover distinctive features on the faces in the lineups. It is possible that having been instructed differently, participants would treat pixelated distinctive features differently. Therefore, we should be careful with the interpretation and generalization of this result.

How does the HS model offer an interpretation of the results of Studies 2 and 3? The interpretation was given on the basis of the differences in the absolute differences of the target:foil familiarity ratios. Under both techniques, familiarity of the target is higher than familiarity of the foils. However, replication amplifies the difference in the familiarity of the target and the foils because of a multiplicative boost in similarity caused by the common distinctive feature, whereas concealment attenuates this difference. Therefore, target selections are more likely under replication.

So far, all of the studies provide evidence for a clear superiority of the replication technique; when the target has a distinctive feature, the distinctive feature should be retained and not concealed at the time of the identification. The next obvious step was to conduct a real-world experiment to test if this result could be replicated in a more ecologically valid paradigm. Although ecological validity is very rarely perfect, the door-to-door experiment of Study 3 was carefully designed, making sure that the correspondence between laboratory- and real-world conditions was very high. First, the sample did not solely consist of university students and second, eyewitnesses, just like in most real-life cases, were not informed that an event would follow during which they would have to encode a face for later recall. However, there was no emotional content in the present experiment (e.g., no increased arousal due to violence) but note that not all real-life identification cases involve witnessing violent crimes. Sometimes eyewitnesses view the perpetrator without being aware that he is about to or has just committed a criminal act. It is true though that we cannot infer that the present results will be generalised to cases involving heightened levels of arousal. The results of this experiment were significant and in line with the previous studies showing a strong effect of the lineup technique on identification accuracy. The initial hypothesis that replication would be a better technique was confirmed.

In this door-to-door experiment, participants were significantly better at identifying the target when the target was present under replication compared to concealment, replicating earlier findings. An additional, unexpected result was revealed: participants were also better at rejecting the lineup when the target was absent under replication compared to concealment. Overall, the replication lineup was more sensitive and at the same time fairer than the pixelation lineup. This

unexpected result was attributed to methodological differences with the previous experiments.

In this experiment, the target was viewed live in participants' office, and he even interacted with them. This may have encouraged people to make abstract inferences (e.g., is he honest?), which has been shown to encourage holistic encoding of faces (Farah, Wilson, Drain, & Tanaka, 1998). Also, the interval between study and test was longer than in all other studies and, most importantly, participants did not expect a memory test. Any of these differentiating factors could be responsible for the observed increased performance under replication in target-absent lineups.

The last experiment (Study 4) failed to replicate the results from all previous experiments using a videotaped event, although the target was the same target as in the real-world experiment. However, this video experiment is less ecologically valid than the door-to-door experiment where the event was experienced incidentally. In Study 4 participants intentionally encoded the target and this might have increased overall identification performance. Indeed, target-identification rates were higher in target-present lineups and foil identifications were lower in target-absent lineups. Also, the time interval between study and test was much shorter (5 minutes instead of 24 hours). Another reason for the increase in identification performance is the fact that participants might have encoded the whole event more meaningfully compared to the previous experiments because they were informed that the study was about how people interpret events and not about memory. Research has shown that deeper, more meaningful processing leads to better memory retention (Lockhart & Craik, 1990). Although part of the answer concerning the contradictory findings may be found in these differences in methodology, again a complete interpretation of the results could not be provided.

Theoretical and Practical Implications

The area of psychology and law –possibly with the exception of the juror decision-making– has long been accused of its inability to contribute to psychological theory (e.g., Johnson, 1993). The present study used a computational model of the effect of distinctive features that had only been applied to recognition-memory experiments, to test if it can be applied in the area of real-world lineups. The HS model was not developed especially to account for data in the area of lineup identifications; however, it appeared successful to model the data of both the face-recognition tasks and the lineup identification tasks, which previous models (e.g., Valentine & Ferrara, 1991) could not account for.

However, it should be recognized that what the HS model describes may not be the only explanation of the outcomes observed in the present studies. Future research that will explicitly aim at testing and refining other models so as to model eyewitness decision-making is strongly encouraged.

From a policy standpoint, the implications from this research are obvious. The number of mistaken identifications in both US and UK are especially alarming. This research indicates that a substantial minority of these people might have been innocent suspects who stood out in the lineup because of their distinctive feature or because of a less fair lineup technique that was used by the police. This research clearly indicates that when suspects have some sort of distinctive facial feature, the distinctive feature should be replicated across lineup members.

The superiority of the replication over the concealment technique is not as intuitive as it may seem with hindsight. There were actually reasons to believe that replication would elicit more mistaken identifications than concealment. Wogalter, Marwitz, and Leonard (1992) manipulated the extent to which the foils resembled

one another and asked people who had never seen the target to pick the most salient person from a lineup. Their findings indicate that high similarity between the foils draws attention towards rather than away from the suspect. This phenomenon is called *similarity bias effect*. Therefore, we would expect that replication –a process in which foils are digitally altered to appear more similar to the suspect– may bias eyewitnesses to select the suspect even when he is not actually the culprit. In the present research, although in target-absent lineups there was no difference in foil-identification rates between replication and concealment, note that target-absent lineups did not include a nominated innocent suspect. This means that in the present experiments, increasing foil similarity could not direct participants' attention to a suspect. Faces were randomly assigned to a role (target or foil) for each participant and all foils were equally similar to one another.

The increase of target-identification rates in target-present lineups under replication is a very important result in itself. Yet target-identification rates are often overlooked in the eyewitness literature because researchers place more emphasis on the consequences of foil-identification rates (i.e., wrongful convictions). However, the consequences of a no-identification (i.e., missing the target) can be of equally high cost given that a free offender might use his freedom to offend again.

Limitations

The present findings are an important first step towards understanding the role of the distinctive features in recognition memory and identification performance, covering to a large extent a huge gap in the eyewitness literature. However, there are four limitations that should be considered. First, note that in all the experiments, foils were chosen based on their similarity to the suspect rather than on the description of the eyewitness. As Luus and Wells (1991) argued the latter technique is more

beneficial to identification performance, a recommendation that was accepted by the Guide. In the present experiments, personal criteria of similarity were used while selecting foils; participants' descriptions of the target were not obtained in any of the experiments. This might have led participants to completely disregard (some of) the foils because they might have been using their own alternate criteria. When foils are chosen based on the eyewitness's description, this is less likely to happen.

Alternatively, the faces that were chosen might have been too similar for the experimental participants, making the identification task rather difficult. However, in an effort to ensure that the observed findings were not an artifact of inappropriate stimulus selection, faces were randomly assigned to each participant. Also, in all the lineup-identification experiments, the different types of lineups were consisted of the same set of foils which means that if participants found the foils too similar under replication, they would also find them too similar under concealment. This ensured that any differences observed would be due to the effects of the presence or absence of distinctive features and not because of differences in other facial components of the foils used.

A second area of concern is the confidence ratings that were obtained in Studies 3 and 4. Pre- (Study 3) as well as post-lineup (Studies 3 and 4) confidence ratings were good predictors of accuracy equally for replication and pixelation/removal lineups, despite the fact that participants were more confident after viewing a replication- rather than a pixelation-technique lineup (Study 3). In line with the initial prediction, the increased confidence ratings under replication were observed only in TP lineups. In TA lineups, post-lineup confidence ratings were similar for replication and pixelation/removal in both Studies.

However, one should be very cautious with the interpretations of these data. The reason is that all of the responses under replication were compared with all of the responses under pixelation and removal. However, many of the participants gave a very low post-lineup confidence rating (lower than 50), which means that the differences in identification performance under replication and under pixelation/removal might have resulted due to different proportions of participants guessing (gave a confidence rating lower than 50). However, the legal system eliminates eyewitnesses who are just guessing, therefore, to generalize the effect of lineup technique on identification accuracy, we should exclude from our analysis those participants whose post-lineup confidence was lower than 50 and examine the difference among the different lineup techniques including the participants who gave post-lineup confidence ratings higher than 50. However, this analysis was not possible in these experiments because of inadequate sample size. Future research examining the effects of lineup techniques on eyewitness confidence and accuracy is strongly encouraged.

A third limitation that must be noted is the use of University students and staff as eyewitnesses. Obviously, participants were not real crime witnesses or victims but experimental participants, hence they were aware that their decision would have no impact on a person's life. We do not know with certainty whether real eyewitnesses would be less or more willing to choose from a lineup. There are reasons to believe it could be either way. On one hand, participants act under the same rules with real eyewitnesses to an extent; they feel they want to please the experimenter by choosing someone and getting it right in the same way that real eyewitnesses feel compelled to "please" the police by being good eyewitnesses who do make a choice. Possibly, the fact that participants were aware that there would be

no consequences following their decision, made a choice even more likely. On the other hand, sometimes eyewitnesses are very motivated to identify the offender, so real eyewitnesses might choose more often. The question whether the results of the University population can be generalized to real eyewitnesses cannot be answered with certainty.

Finally, a limitation that applies to all lineup-identification studies of this thesis is the replication of the identical distinctive feature of the culprit across lineup members in target-absent lineups. Although this technique is perfectly appropriate for those cases where the distinctive feature is one that is frequently encountered on faces (e.g., a mole) and has been accurately described by the eyewitness (e.g., the exact location on the face, the size etc.), it is less plausible for those cases where the suspect has a unique distinctive feature (e.g., a web tattoo). Unless the eyewitness picks the exact tattoo pattern from a large database of tattoo patterns, we have no way of knowing what the offender's tattoo looked like and therefore we are unable to replicate it. The cases in which an innocent suspect has exactly the same distinctive feature as the offender are extremely unlikely. Hence we might predict that if the distinctive features in target-absent lineups were new, unseen distinctive features, identification accuracy under replication would be even higher because new faces with unseen distinctive features would be more likely to be rejected as unseen faces than faces with a distinctive feature that was previously seen. This issue, among other areas for future research, are discussed below.

Future Directions

I close this thesis by suggesting seven directions for future experimental work in the area of lineup construction for suspects with distinctive features. I also discuss what other methods might be fruitfully considered for future use.

First, one interesting direction for future research would be to manipulate the extent to which the distinctive feature of the suspect matches the distinctive features of the other lineup members. In the current research, replication was applied by replicating the identical distinctive feature of the target across all lineup members. Such a technique might have made the lineup members to resemble one another at a high level, making the identification task rather difficult for the participants. Researchers (e.g., Turtle et al., 2003; Valentine, 2006; Wells, Rydell, & Seelau, 1993) argue that a lineup should be fair but at the same time sufficiently sensitive to allow a witness who has a good memory of the culprit to make a positive identification. They argue that while constructing lineups, care should be taken so as to increase sensitivity without increasing fairness.

Following with this logic, in the case of a suspect with a distinctive feature, a lineup with only one person (i.e., the suspect) having the distinctive feature that the witness mentioned would increase sensitivity but at the same time it would decrease fairness. Equally, a lineup where everyone has the suspect's distinctive feature would increase fairness but at the expense of sensitivity; even a reliable eyewitness would have a low chance of making a positive identification.

Based on this logic, Valentine, Hughes, and Munro (2009) suggested that the use of a technique of "replication with variation" rather than replication of the identical distinctive feature on each lineup member would increase lineup sensitivity without reducing lineup fairness. It is important though that the variation of the replicated feature be based on the eyewitness's description. For example, if the witness referred to a tattoo of a star on the culprit's forehead, the colour, location, and size of the tattoo should vary among lineup members. Valentine et al.'s rationale is that if the culprit is in the lineup, eyewitnesses can use their memory of the tattoo

to identify the culprit. However, variation will not bias the lineup against an innocent suspect who an eyewitness has not seen before.

In general, the purpose of a lineup is to constitute an extra test of memory, beyond the eyewitness's description. If, for example, the eyewitness has given details about a star tattoo on a specific location of the culprit's face but in the lineup, the location, size, and colour of the tattoo varies among lineup members, then the choice of the eyewitness tells the police nothing new over and above the eyewitness's initial description. The police might have as well just comprehended the suspect only on the basis of the description. In other words, in the cases where the distinctive feature is common and we replicate with variation, target identifications might be increased in target-present lineups (for the wrong reasons), but foil identifications might also be increased in case the innocent suspect has a distinctive feature very similar to the culprit's one. Sensitivity is then increased at the expense of fairness.

To conclude, replication should be applied based on the degree of detail that exists in the eyewitness's description. If the description of the distinctive feature is detailed, then the purpose of the lineup would be to identify the right face, and so, identical distinctive features should be applied. If the description of the distinctive feature is vague, then the purpose of the lineup would be to identify the right distinctive feature, and therefore, replicating with variation is the most appropriate technique. Of course there are cases that it is not as clear which technique must be applied. In these cases common sense should prevail. The basic rationale behind the decision must be that the lineup constitutes an extra test of memory beyond the eyewitness's description and that it should either target memory for the distinctive feature, or memory for the face.

Second, a direction for future research concerns the investigation of the effect of each lineup technique on accuracy when the eyewitness's description does not refer to the suspect's distinctive feature. For reasons that were covered in Study 3, some eyewitness descriptions of the culprit do not refer to a distinctive feature of the suspect. In such cases, the Guide recommends that the relevant area on the face of all lineup members should be concealed. However, there are reasons to believe that it may be better to replicate the distinctive feature on all lineup members. Research has shown that people are able to encode information (and recall it in an implicit test, e.g., completion) that they do not recall in their description (Shanks & St. John, 1994). Therefore witnesses may be able to recognise a distinctive feature should they see it on the face of the culprit in a lineup, despite not being able to verbalize the presence of the feature. This is something that needs to be investigated.

Third, an interesting future direction would be to investigate the effects of suggestive instructions during the administration of a lineup. When a distinctive feature has been covered with a mask, current lineup instructions inform eyewitnesses that, in order to make the lineup procedure fair, there are masks on the images they are about to see which may or may not conceal features on the faces of those individuals (Karl Burns, Northumbria Identification Unit, personal communication, January 30, 2009). However, eyewitnesses are informed that there exists another photo of the same people but without any sort of concealment on their faces. Eyewitnesses are instructed further that if, after viewing the lineup at least twice, they would like to see *one* lineup member without that concealment, then they are allowed to do so by telling the lineup administrator the number that corresponds to this particular lineup member. However, from a psychological point of view, this is a highly suggestive procedure which may imply that the culprit is in the lineup.

Such an instruction encourages the eyewitness to select someone from the lineup rather than reject the lineup. Also, the presentation of the photo of one lineup member without the concealment does not provide any further test of the eyewitness' memory.

Fourth, as mentioned in the introductory section of this thesis, video-lineups have replaced live lineups because this format has proved to be fairer and it is preferred for identification procedures in the UK. However, replication software for moving images is costly to use daily and police officers use still images when replication is applied. In further experiments, "motion with replication" can be achieved with transfer tattoos on the faces of the foils and be compared to still, concealed images to see if motion with replication is significantly superior to still images and should be preferred despite its high cost.

Fifth, in the current research only simultaneous lineups were used, that is, all lineup faces were presented simultaneously to participants. However, Lindsay and Wells (1985) have proposed that lineup faces should be presented one at a time, sequentially. As discussed in the introductory chapter of this thesis, during a sequential procedure, the eyewitness is instructed that they are going to be viewing one face at a time and have to decide for each face whether it is the culprit or not, without knowing the total number of the lineup-faces. In this way, eyewitnesses are asked to make absolute- instead of relative-judgment decisions that are thought to lead to more identification errors (Gronlund, 2004). This new method has been adopted by police officers who often present lineups sequentially to eyewitnesses, although the debate about the superiority of one presentation mode over the other is still ongoing. The results from the current research should generalise to sequential lineups as long as eyewitnesses are informed that the lineup members have replicated

distinctive features. This issue should be explored in future research, although, as both modes have been proved to be equally problematic (Levi, 1998), many researchers might want to follow Levi's (1998) suggestion to eyewitness scientists to direct their efforts towards new methods of identification tests.

Sixth, it would also be of interest to examine the effect of the location of the distinctive feature on the face of the culprit. As discussed earlier in this thesis, in order to recognize unfamiliar faces, people tend to rely more on external features, such as the hair and the face outline, whereas for the recognition of familiar faces, they tend to rely more on internal features such as the eyes, the nose and the mouth (Bonner, Burton, & Bruce, 2003). One can hypothesize that distinctive features located near the periphery of the face may be more likely to lead participants to encode facial components that are crucial to recognition of unfamiliar faces, like the hair line, the face outline etc. Similarly, distinctive features located near the centre of the face should be more likely to impair recognition performance as in this case, participants' attention will be attracted far from the periphery of the face. It should be noted here, that although the eyes are considered an internal feature, they have been proved to be detrimental to recognition (O'Donnell & Bruce, 2001) because we tend to selectively attend to the eyes more during communication (e.g., Langton & Bruce, 1999). For this reason, concealment of a distinctive feature in the eye area may lead to more mistaken identifications.

Finally, future research on the effects of distinctive features should take advantage of the methodological advances in techniques that allow more objective measurements of these effects. For example, one of the best methods to investigate whether eyewitness' attention is shifted towards the distinctive feature of a culprit at the time of encoding is the eye-tracking technique. This technique was used initially

in the area early on in the history of the discipline by Loftus, Loftus, and Messo (1987) to reveal eyewitnesses' tendency to focus disproportionately on the weapon of the culprit. Eye-tracking during test, would reveal whether eyewitnesses apply relative judgment by comparing the distinctive features or other facial characteristics. This methodology would be very informative. Also, field experiments with actors as the culprits of staged events after which witnesses will view a lineup using one of the techniques including both target-present and target-absent lineups must be employed to investigate the effect of the techniques on the probability of correctly identifying a guilty suspect and the potential for mistaken identification of an innocent suspect.

In closing, it is important to note that the design of all the above empirical studies that have been suggested should be informed by a survey of current police practice for constructing lineups for suspects with distinctive features. The survey would indicate how this research might inform police practice. The survey should be aimed at collecting records of cases, the lineup technique that was used for each case, as well as the eyewitness's identification decision after viewing the lineup. What do the police consider to be a distinctive feature? In other words, what are the specific characteristics of a feature –deemed by the police to be distinctive– that lead the police to apply a special procedure?

Given the lack of empirical research concerning lineup techniques for suspects with distinctive features, I believe it is premature to offer recommendations for policy. Although the results of this thesis are pointing towards a specific direction, we should be very cautious with their interpretations and their potential generalizability to real-world cases.

Conclusion

The goal of this research was to explain how two different methods of constructing lineups for suspects with distinctive features lead to specific types of identification errors. Although much research exploring these techniques is still to be done, the findings of the current investigation suggest that replicating the distinctive feature across lineup members is a preferred technique for constructing fair lineups for suspects with distinctive features compared to concealing the distinctive feature on the face of the suspect. It is hoped that eyewitness scientists will embrace this finding to develop a body of theoretical work and practitioners and policy makers will be aware of this potential effect of lineup technique on identification performance when constructing lineups for suspects with distinctive features.

References

- Bartlett, J., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*, 219-228.
- Birnbaum, M. H. (Ed.). (2000). *Psychology experiments on the internet*. San Diego, CA: Academic Press.
- Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition*, *10*, 527-536.
- Bradfield, A. L., & Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. *Law and Human Behavior*, *24*, 581-594.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behaviour Therapy and Experimental Psychiatry*, *25*, 49-59.
- Brewer, N., Weber, N., & Semmler, C. (2005). Eyewitness identification. In N. Brewer & K. D. Williams (Eds.), *Psychology and law: An empirical perspective* (pp. 177-221). New York: Guilford.
- Brigham, J. C., Ready, D. J., & Spier, S. A. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology*, *11*, 149-163.
- Bruce, V. (1982). Changing faces: Visual and non-visual coding processing in face recognition. *British Journal of Psychology*, *73*, 105-116.
- Bruce, V., Burton, A. M., & Dench, N. (1994). What's distinctive about a distinctive face? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *47*, 119-141.

- Busey, T. A. (2001). Formal models of familiarity and memorability in face recognition. In: M. J. Wenger & J. T. Townsend (Eds.) *Computational, Geometric, and Process Perspectives on Facial Cognition*. New Jersey: Erlbaum Associates.
- Charman, S. D., & Wells, G. L. (2007). Eyewitness lineups: Is the appearance-changes instruction a good idea? *Law and human behavior*, *31*, 3-22.
- Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, *17*, 629-654.
- Cutler, B. L. & Penrod S. D. (1989). Forensically-relevant moderators of the relationship between eyewitness identification accuracy and confidence. *Journal of Applied Psychology*, *74*, 650-652.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987a). The reliability of eyewitness identification: The role of system and estimator variables. *Law and Human Behavior*, *11*, 233-258.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987b). Improving the reliability of eyewitness identification: Putting context into context. *Journal of Applied Psychology*, *72*, 629-637.
- Cutler, B. L., Penrod, S. D., O'Rourke, T. E., & Martens, R. K. (1986). Unconfounding the effect of contextual cues on eyewitness identification accuracy. *Social Behavior*, *2*, 113-134.
- Darling, S., Valentine, T., & Memon, A. (2008). Selection of lineup foils in operational contexts. *Applied Cognitive Psychology*, *22*, 159-169.
- Deffenbacher, K. A., Johanson, J., Vetter, T., & O'Toole, A. J. (2000). The face typicality-recognizability relationship: Encoding or retrieval locus? *Memory & Cognition*, *28*, 1173-1182.

- Dennis, I. (2007). *The law of evidence*. Sweet & Maxwell: London.
- Dodson, C. S. & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, 8, 155-161.
- Doob, A. N., & Kirshenbaum, H. (1973). Bias in police lineups -- partial remembering. *Journal of Police Science and Administration*, 1, 287-293.
- Ellis, H. D., Davies, G. M., & Shepherd, J. W. (1977). Experimental studies of face identification. *Journal of Criminal Defence*, 3, 219-234.
- Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is 'special' about face perception? *Psychological Review*, 105, 482-498.
- Frumkin, D., Wasserstrom, A., Davidson, A., & Grafit, A. (in press). Authentication of forensic DNA samples. *Forensic Science International: Genetics*.
- Gonzalez, R., Ellsworth, P. C., & Pembroke, M. (1993). Response biases in lineups and showups. *Journal of Personality and Social Psychology*, 64, 527-537.
- Gronlund, S. D. (2004). Sequential lineups: shift in criterion or decision strategy? *Journal of Applied Psychology*, 89, 362-368.
- Home Office (2008). Police and Criminal Evidence Act 1984 (PACE) and accompanying Codes of Practice. Retrieved August 16, 2009 from <http://police.homeoffice.gov.uk/operational-policing/powers-pace-codes/pace-codes.html>.
- Innocence Project (n.d.). Retrieved August 16, 2009, from <http://www.innocenceproject.org/understand/Eyewitness-Misidentification.php>
- Iowa State University (n.d.). Retrieved September 28, 2009, from <http://www.psychology.iastate.edu/~glwells/homepage.htm>

- Johnson, M. T. (1993). Memory phenomena in the law. *Applied Cognitive Psychology, 7*, 603-618.
- Johnston, W. A., Hawley, K. J., Plewe, S. H., Elliott, J. M. G., & DeWitt, M. J. (1990). Attention capture by novel stimuli. *Journal of Experimental Psychology: General, 119*, 397-411.
- Jones, T. C., Bartlett, J. C., & Wade, K. A. (2006). Nonverbal conjunction errors in recognition memory: Support for familiarity but not for feature bundling. *Journal of Memory and Language, 55*, 138-155.
- Kassin, S. M., & Neumann, K. (1997). On the power of confession evidence: An experimental test of the fundamental difference hypothesis. *Law & Human Behavior, 21*, 469-484.
- Knapp, B. R., Nosofsky, R. M., & Busey, T. A. (2006). Recognizing distinctive faces: A hybrid-similarity exemplar model account. *Memory and Cognition, 34*, 877-889.
- Langton, S. R. H., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition, 6*, 541-567.
- Laughery, K., Alexander, J., & Lane, A. (1971). Recognition of human faces: Effects of target exposure, target position, pose position, and type of photograph. *Journal of Applied Psychology, 55*, 477-483.
- Levi, A. M. (1998). Are defendants guilty if they were chosen in a lineup? *Law & Human Behavior, 22*, 389-407.
- Levi, A. M. (2006). An analysis of multiple choices in MSL lineups, and a comparison with simultaneous and sequential ones. *Psychology, Crime, & Law, 12*, 273-286.

- Levi, A. M. (2007). Evidence for moving to an 84-person photo lineup. *Journal of Experimental Criminology, 3*, 377-391.
- Levi, A. M., & Lindsay, R. C. L. (2001). Lineup and photospread procedures: Issues concerning policy recommendations. *Psychology, Public Policy, & Law, 7*, 776-790.
- Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 212-228.
- Lindsay, R. C. L. & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law & Human Behavior, 4*, 303-313.
- Lindsay, R. C. L. & Wells, G. L. (1985). Improving eyewitness identification from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*, 556-564.
- Lockhart, R. S., & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology, 44*, 87-112.
- Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about "weapon focus". *Law & Human Behavior, 11*, 55-62.
- Luus, C. A., & Wells, G. L. (1991). Eyewitness identification and the selection of distractors for lineups. *Law and Human Behavior, 15*, 43-57.
- Malpass, R. S. (2006). A policy evaluation of simultaneous and sequential lineups. *Psychology, Public Policy, & Law, 12*, 394-418.

- Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, 66, 482-489.
- Malpass, R. S., & Devine, P. G. (1981). Guided memory in eyewitness identification. *Journal of Applied Psychology*, 66, 343-350.
- Memon, A., & Gabbert, F. (2003). Unravelling the effects of sequential presentation in culprit-present lineups. *Applied Cognitive Psychology*, 17, 703-714.
- Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. *British Journal of Psychology*, 94, 339-354.
- Murdock, B. B. (1960). The distinctiveness of stimuli. *Psychological Review*, 67, 16-31.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Nosofsky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1194-1209.
- O'Donnell, C., & Bruce, V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*, 30, 755-764.
- O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and other-race effect. *Memory & Cognition*, 22, 208-224.
- Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology. Human Learning and Memory*, 3, 406-417.

- Read, J. D. (1995). The availability heuristic in person identification – the sometimes misleading consequences of enhanced contextual information. *Applied Cognitive Psychology, 9*, 91-121.
- Read, J. D., Vokey, J. R., & Hammersley, R. (1990). Changing photos of faces: Effects of exposure duration and photo similarity on recognition and the accuracy-confidence relationship. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 870-882.
- Scheck, B., Neufeld, P., & Dwyer, J. (2001). *Actual innocence*. New York: Penguin Putnam.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.
- Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367-447.
- Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*, 139-156.
- Shepherd, J. W., Gibling, F., & Ellis, H. D. (1991). The effects of distinctiveness, presentation time and delay on face memory. *European Journal of Cognitive Psychology, 3*, 137-145.
- Snodgrass, J. G., Corwin, J. (1988). Perceptual identification thresholds for 150 fragmented pictures from the Snodgrass and Vanderwart picture set. *Perceptual & Motor Skills, 67*, 3-36.
- Stebay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law & Human Behavior, 25*, 459-473.

- Technical Working Group on Eyewitness Evidence. (1999). *Eyewitness Evidence: A Guide to for law enforcement*. Washington DC: US Department of Justice http://www.ncjrs.org/nij/eyewitness/tech_working_group.html [18/3/2009].
- Tredoux, C. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. *Journal of Experimental Psychology: Applied*, 8, 180-193.
- Tredoux, C. G., Meissner, C. A., Malpass, R. S., & Zimmerman, L. A. (2004). Eyewitness identification. In C. Spielberger, s (Ed.), *Encyclopedia of Applied Psychology* (pp. 875-887). San Diego, CA: Academic Press.
- Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, 20, 479–496.
- Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352-373.
- Turtle, J. W., Lindsay, R. C. L., & Wells, G. L. (2003). Best practice recommendations for eyewitness evidence procedures: New ideas for the oldest way to solve a case. *The Canadian Journal of Police and Security Services*, 1, 5-18.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.
- Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43, 161-204.
- Valentine, T. (2001). Face-space models of face recognition. In: M. J. Wenger & J. T. Townsend (eds.) *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*. Mahwah: LEA (pp.83-113).

- Valentine, T. (2006). Forensic facial identification. In: Heaton-Armstrong, A., Shepherd, E., Gudjonsson, G. & Wolchover, D. (eds). *Witness Testimony; Psychological, Investigative and Evidential Perspectives* (pp. 281-307). Oxford: Oxford University Press.
- Valentine, T. & Bruce, V. (1986a). The effects of distinctiveness in recognizing and classifying faces. *Perception, 15*, 525-535.
- Valentine, T. & Bruce, V. (1986b). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology, 40*, 300-305.
- Valentine, T. & Ferrara, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology, 82*, 87-102.
- Valentine, T., & Heaton, P. (1999). An evaluation of the fairness of police lineups and video identifications. *Applied Cognitive Psychology, 13*, 59-72.
- Valentine, T., Darling, S., & Memon, A. (2007). Do strict rules and moving images increase the reliability of sequential identification procedures? *Applied Cognitive Psychology, 21*, 933-949.
- Valentine, T., Harris, N., Colom Piera, A., & Darling, S. (2003). Are police video identifications fair to African-Caribbean suspects? *Applied Cognitive Psychology, 17*, 459-476.
- Valentine, T., Hughes, C. & Munro, R. (2009). Recent developments in eyewitness identification procedures in the United Kingdom. In: Bull, R., Valentine, T. & Williamson, T. (eds.) *The handbook of psychology of investigative interviewing*. Chichester: Wiley. (pp.221-240).

- Valentine, T., Pickering, A., & Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology, 17*, 969-993.
- Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*, 291-302.
- Wells, G. L. & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness non-identifications. *Psychological Bulletin, 88*, 776-784.
- Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 12*, 1546-1557.
- Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*, 89-103.
- Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553-571.
- Wells, G. L. (1995). Scientific study of witness memory: Implications for public policy and law. *Psychology, Public Policy, & Law, 1*, 726-731.
- Wells, G. L., & Bradfield, A. L. (1998). 'Good, you identified the suspect': Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376.
- Wells, G. L., & Olson, E. (2003). Eyewitness identification. *Annual Review of Psychology, 54*, 277-295.
- Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*, 835-844.

- Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law & Human Behavior, 22*, 603-647.
- Wells, G.L., Malpass, R.S., Lindsay, R.C.L., Fisher, R.P., Turtle, J.W., & Fulero, S. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist, 55*, 581-598.
- Wogalter, M. S., Malpass, R. S., & McQuiston, D. E. (2004). A national survey of U.S. police on preparation and conduct of identification lineups. *Psychology, Crime, & Law, 10*, 69–82.
- Wogalter, M. S., Marwitz, D. B., & Leonard, D. C. (1992). Suggestiveness in photospread lineups: Similarity induces distinctiveness. *Applied Cognitive Psychology, 6*, 443-453.
- Wixted, J. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 681-690.
- Zarkadi, T., Stewart, N., & Wade, K. A. (2009). *Lineup construction for suspects with distinctive features: To replicate, pixelate, or remove?* Manuscript submitted for publication.
- Zarkadi, T., Wade, K. A., & Stewart, N. (in press). Creating fair lineups for suspects with distinctive features. *Psychological Science*.