Evaluating the fairness of identification parades with measures of facial similarity.

A thesis, submitted for the degree of doctor of Philosophy (Ph.D.), in the faculty of Social Sciences and Humanities, University of Cape Town,

by

Colin Getty Tredoux (M.A., UCT)

Supervisor: **Professor Don Foster**
Department of Psychology
University of Cape Town

February, 1996

27 JAN 1897

BUT 150 TRED
96/11 184

Acknowledgements

The work reported in this (blessed) volume has taken me the better part of five years. Many have suffered while it trundled along, inexorably. I am grateful for their sufferance, and more grateful still for their support. I owe much to many.

Don Foster, my supervisor, has guided wisely, from near and far. He is aptly recognised as the 'doyen of social psychology in South Africa'.

My colleagues and friends from social psychology circles, John Dixon, Kevin Durrheim, Don Foster, and more recently Gillian Finchilescu, Cheryl de la Rey, Pumla Gobodo, Neo Morojele, have goaded, chided and heartened me to this task.

Other colleagues and friends, in many places, have played important roles, too, in the many ways in which such people do. Peter du Preez, Andy Dawes, Johann Louw, Michael, Pluto, Damon, Roy, Armien, Heather, the many tutors I have worked with in undergraduate programs at U.C.T., my legal conscriptors Paul and Shereen, the department of Psychology at UCT - there are many to thank. My mother also contributed, just when she thought such contributions had ceased.

My wife, Diane Gascoigne, has encouraged, cajoled, commiserated, contained. A fixed mark, a constant star in a sea of change.

I acknowledge also the financial assistance of the Centre for Science Development (*HSRC*, South Africa) towards this thesis. Opinions expressed in the work, or conclusions arrived at, are mine, and should not be attributed to the Centre for Science Development.

Abstract

This thesis addresses a practical problem. The problem concerns the evaluation of 'identification parades', or 'lineups', which are frequently used by police to secure evidence of identification. It is well recognised that this evidence is frequently unreliable, and has led on occasion to tragic miscarriages of justice. A review of South African law is conducted and reported in the thesis, and shows that the legal treatment of identification parades centres on the requirement that parades should be composed of people of similar appearance to the suspect. I argue that it is not possible, in practice, to assess whether this requirement has been met and that this is a significant failing. Psychological work on identification parades includes the development of measures of parade fairness, and the investigation of alternate lineup structures. Measures of parade fairness suggested in the literature are indirectly derived, though, and I argue that they fail to address the question of physical similarity. In addition, I develop ways of reasoning inferentially (statistically) with measures of parade fairness, and suggest a new measure of parade fairness.

The absence of a direct measure of similarity constitutes the rationale for the empirical component of the thesis. I propose a measure of facial similarity, in which the similarity of two faces is defined as the Euclidean distance between them in a principal component space, or representational basis. (The space is determined by treating a set of digitzed faces as numerical vectors, and by submitting these to principal component analysis). A similar definition is provided for 'facial distinctiveness', namely as the distance of a face from the origin or centroid of the space.

The validity of the proposed similarity measure is investigated in several ways, in a total of seven studies, involving approximately 700 subjects. 350 frontal face images and 280 profile face images were collected for use as experimental materials, and as the source for the component space underlying the similarity measure. The weight of the evidence, particularly from a set of similarity rating tasks, suggests that the measure corresponds reasonably well to perceptions of facial similarity. Results from a mock witness experiment showed that it is also strongly, and monotonically related to standard measures of lineup fairness. Evidence from several investigations of the distinctiveness measure, on the other hand, showed that it does not appear to be related to perceptions of facial distinctiveness. An additional empirical investigation examined the relation between target-foil similarity and identification performance. Performance was greater for lineups of low similarity, both when the perpetrator was present, and when the perpetrator was absent. The consequences of this for the understanding of lineup construction and evaluation are discussed.

Contents

Tables

Figures

Cases

Chapter 1 — Introduction

This thesis addresses a practical problem. The problem concerns the evaluation of an identification technique commonly used by police, which is known as the 'identification parade', or 'lineup'.[1] The quality of evidence secured from an identification parade is dependent on both the structure of the parade, and the way in which it is regulated. Legal systems in most countries have evolved fairly effective conditions for the regulation of parades, but have paid little attention to the question of parade structure, particularly with respect to how parade structure should be evaluated. The central requirement of parade structure, at least in English and South African law, is that it consist of an adequate number of foils who bear a sufficient physical resemblance to the suspect. The courts provide ineffective criteria for judging 'sufficient physical resemblance', and this hampers the task of parade construction, and the task of parade evaluation. Since the identification parade is considered a key test of eyewitness identification, it is clear that the inability to assess the requirement of physical similarity is a significant problem. The central project of this thesis is the development and validation of a measure of one element of physical similarity. In particular, I will propose a measure of facial similarity, which appears also to be a promising measure of parade fairness. This thesis is consequently a treatise in 'applied psychology', but not narrowly so. There are important issues, in particular, that attend the notion of 'applied psychology', and a consideration of these will take us well outside the project of the similarity measure.

Psychology has long suffered an internal division over its site of practice. Experimental psychology has, by and large, opted for the artificial laboratory, in wilful emulation of natural sciences like Chemistry. At the same time, other traditions of psychology have chosen locations outside the confines of the laboratory, in the world of 'practical problems'. The present piece of work adopts the approach of the latter type of tradition, but will endeavour also to examine its foundations. There is no attempt to exalt this approach over others, but the inherent jealousy of intellectual traditions should, of course, be recognised.

One of the 'real problems' that Psychology has addressed on several occasions this century, but especially in recent times, is that of legal identification evidence. Humans who witness events are

---

[1] The terms 'identification parade' and 'lineup' will be used interchangeably in this thesis.

frequently called by courts and police services to deliver their testimony. This testimony frequently takes the form of an identification of a person allegedly involved in the witnessed event, and is used as evidence against the identified person. It is well known that eyewitness identifications are frequently mistaken, and have on occasions led to tragic miscarriages of justice. Several of the more recent tragedies in the USA are vividly documented by Loftus & Ketcham (1991).[2] Some legal commentators suggest that mistaken identifications are the single greatest cause of legal injustice,[3] and courts have instituted several safeguards against them. The safeguard perhaps held in highest regard is a technique known as the 'identification parade'. A person suspected of a crime is made to stand alongside a number of other people, of similar height, build, and physical appearance. The witness is asked to point out the perpetrator of the crime from the array of assembled people. An identification from a 'regularly conducted' parade is given great weight, particularly in the criminal justice system of South Africa, which is the legal frame of reference for this study.

Identification parades, however, are not a guarantee of the veracity of witness identifications. Indeed, a series of mistaken parade identifications in the celebrated case of Adolf Beck probably brought about the institution of the English Court of Criminal Appeal (Shepherd, Ellis & Davies, 1982). Legal systems all over the world are mindful that identification parades do not provide the guarantee supposed of them: There have been no fewer than three commissions of enquiry into identification evidence in England, alone, this century (Devlin, 1976).

Psychologists have applied themselves to the problem of identification parades in recent times, and their contributions include the development of measures of parade fairness, and the invention of alternate parade structures. This is a kind of 'applied psychology', an empirical search for solutions to the problems that identification parades pose. But the notion of an 'applied psychology' is not as uncomplicated as it may first appear. Indeed, I will argue in Chapter 2 that the notion of an 'applied psychology' is typically misunderstood. It is most commonly taken to be 'the application of known methods to real problems', but very little of what goes by the name of applied psychology is ever applied. Most applied psychology is at best 'applicable psychology', and the label 'applied' is used merely by dint of the topic of investigation. In short, there is a sort of nominalist error: because a name exists, an entity is presumed to correspond to it. In this way, 'Applied Psychology' has come

---

[2] South Africa has surely had a number of tragedies induced by mistaken identifications, but there is no systematic documentation to support this presumption. A number of near-tragedies are documented, though. In one of the famous 'crowd-murder' - or 'collective action' - cases of the mid 1980's, known as *Duduza*, one of the accused was identified by several witnesses, and was saved only by video footage which showed that the perpetrator was left-handed, whereas the accused was right-handed. (Tyson, personal communication). A newspaper report, which is reproduced on page 59, also leads one to believe that misidentifications are not uncommon in South Africa.

[3] This is discussed at some length in Chapter 3.

to be treated as a subdiscipline of psychology, few questions are asked about the nature of 'application', and few researchers show how their studies can be used in respect of the very problems that serve as justification for the research in the first place.

In some respects, the psychological literature on identification parades is heedful of these problems. Thus, Gary Wells has written on several occasions about the applied nature of research on identification evidence (Wells, 1978; Wells, 1986), and has outlined strategies that lead or could lead to application of research. His contributions in this respect will be considered closely in Chapter 2. I draw several conclusions in Chapter 2, but the important conclusion taken forward from that chapter is that 'practical problems' require careful analysis, and should not be approached peremptorily. In particular, they require careful analysis in the terms of the context or discipline in which the problem is located. In the present case, this implies that the problems of identification evidence and identification parades should be understood in terms of the referent (i.e. South African) criminal justice system.

Accordingly, Chapter 3 presents a review of South African law on identification evidence. Material from statutes, case law, and internal police documents will be considered. I will try to draw a coherent picture of South African legal practice in respect of identification parades, but there are several contradictions and complications. In brief, South African law has long recognised the treacherous nature of identification evidence, and has devised a system of rules and safeguards against it. The identification parade is probably considered the strongest of these safeguards, but the special dangers that attend this technique are also well recognised. Many requirements and strictures regulate the conduct of the parade in South Africa, but I will argue that these are subsidiary in an important sense to the central requirement that the parade consist of the suspect, and a number of innocent people who are sufficiently similar in appearance to the suspect. Guidelines for evaluating this requirement have been handed down by the courts, but are extremely vague. I will argue that this 'central requirement' cannot be met in practice, either at the level of police regulation of identification parades, or at the level of judicial evaluation.

South Africa is certainly not the only country faced with problems stemming from identification parades. Psychologists in several countries have suggested useful measurement techniques, and have investigated modifications aimed at improving the lineup as an identification procedure. This work is considered at some length in Chapter 4. Although the problem of identification evidence appears to invite 'thoroughly practical' research, the review presented in Chapter 4 will show that there is also some theoretical work of consequence. Two competing Bayesian approaches are outlined: in the formulation due to Navon (1990a, 1990b), identification parades are postulated to provide information about the physical resemblance of perpetrator and suspect, whereas the

alternate formulation due to Wells and colleagues (Wells & Lindsay, 1980; Wells & Luus, 1990a) is said to provide information about the reliability of the parade identification.[4]

I also consider the recent attempt by Wells, Seelau, Rydell & Luus (1994) to formulate the rudiments of a theory of identification parades. Wells et al. suggest two propositions, and a corollary. The propositions are i) that identification parades are recognition tests, which seek to uncover information not present at recall, and ii) that the identification process is governed in part by extramemorial judgement and heuristic processes. The corollary is that a lineup task can be likened to a social psychology experiment: factors that confound such an experiment can also confound the lineup task.

The theoretical work on identification parades is useful, but I argue that a central theoretical tangle remains unresolved. In particular, the nature of the task embodied in the identification parade is not clear: is the task a reliability test, or does it collect independent evidence of identity? I address this uncertainty at several places in the thesis; see especially Chapters 3, 4, and 6.

The groundbreaking studies of Doob & Kirshenbaum (1973), Wells, Leippe & Ostrom (1979), Malpass (1981), and Malpass & Devine (1983) are among the most useful of all psychological work on identification parades. These studies propose methods for measuring the 'fairness' of identification parades, and rely on a method of soliciting 'mock' eyewitness identifications, known as the mock witness task. This task requires 'witnesses' to guess the identity of the perpetrator, even though they have not previously seen the perpetrator, and have only a brief verbal description of him. If 'mock witnesses' are able to guess the identity of the suspect at a level of success greater than would be expected under an equiprobability model, the lineup is said to be biased against the suspect. In addition, the distribution of identifications across lineup foils sustains several other measures of lineup fairness. These concern the effective 'size' of the lineup, which may be defined as the number of foils who are effective choice alternatives: if fewer than the total number of foils attract mock witness identifications, the lineup's effective size is smaller than its nominal size.

The mock witness task assumes a probability model for interpretation, but this is not made explicit in the psycho-legal literature, nor is any advice given regarding the statistical evaluation of mock identifications. This is a significant failing, and I attempt to remedy it in Chapter 6. I argue that a binomial probability model can be used to model the mock witness task, and suggest ways in which inferential statistics can be used to assist in the interpretation of indices of lineup bias and lineup

[4] This description of the Bayesian approach taken by Wells and colleagues is due to Navon (1990a), and is arguable. The dispute between the two approaches is detailed in Chapter 4.

size. Certain additional techniques suggested by Malpass & Devine (1983, 1984) for the selection and evaluation of parade foils, which are based in part on the mock witness method, are considered, and suggestions are made regarding the application of inferential statistical thinking to these techniques. In Chapter 6, I also explore statistical methods for assistance in the interpretation of other lineup indices, including the Bayesian measures of 'diagnosticity', and 'information gain' devised by Wells & Lindsay (1980).

Lineup fairness indicators based on the mock witness task are indirect measures. They show a lineup to be biased against the suspect, if she is chosen at levels greater than chance expectation, or they show the lineup to have low effective size, if certain foils are chosen at levels which deviate from chance expectation. They do not show which aspect of the parade has led to a lack of 'fairness', but this is presumably the absence of a sufficient number of foils who resemble the description give to mock witnesses, and by implication, a similar absence of foils who resemble the suspect.

I argue in the conclusion of Chapter 4 that a direct measure of suspect-foil similarity would be a useful addition. Although a whole-body measure of similarity would be a more complete solution, facial similarity is probably the most important element of physical resemblance,[5] and the major empirical goal of the thesis is to develop a direct way of measuring this.

The search for a measure of facial similarity takes us to Chapter 5, and the face recognition literature. The size of this literature demands a focused approach; my approach in that chapter is therefore a shameless raid of the literature for a suitable conceptualization of facial similarity. I will consider several possibilities, and argue for an adaptation of a particular approach to facial representation: this approach represents faces as eigenspaces of normalized intensity maps in the picture plane, and it allows - I suggest - a method of deriving a similarity space for populations of facial images. The route to this position is not uncomplicated, or brief; it takes us through much recent theoretical and empirical work, and there are several profitable stops.

The influential theoretical models of face recognition advanced by Bruce and colleagues (Bruce & Young, 1986; Burton & Bruce, 1990) are considered, but the models do not explicitly address perceptions of facial similarity. Nevertheless, the intimation of a visual-front end to the models is taken forward, as is the notion of a 'canonical representational code'. Specific treatment of facial similarity can be found in a handful of studies, and the most useful of these for my purposes appear

[5] This is probably the rationale for the common modern police practice of using lineups constituted by head-and-shoulder photographs of the suspect and foils. The widespread use of these 'photo-parades' is justification enough, in my opinion, for the restriction adopted here.

to be those reported by Valentine and colleagues (Valentine 1991a, 1991b; Valentine & Endo, 1992; Valentine and Ferrara, 1991). Valentine suggests a theory of the perceptual representation of faces in a multi-dimensional space, and such a space easily allows the derivation of facial similarity and facial distinctiveness measures as spatial distances, or dot products. Unfortunately, Valentine's model is only formulated at a conceptual level, and he leaves the dimensions that constitute the multidimensional space unspecified. Later sections of the chapter take me in search of these 'dimensions', and I conclude that the most promising solution here is the representational schema developed by Sirovich & Kirby (1987), O'Toole & colleagues (e.g. O'Toole & Thompson, 1993), and earlier, by Kohonen (1984). This scheme is the 'engine' for the facial similarity measure used in the empirical component of the thesis, and is described in some detail. It proposes treating digitized face images as numerical vectors, and finding a basis for the space constituted by these vectors, through principal component analysis.

This solution to the problem of finding a representational basis strikes me as particularly elegant: it allows one to represent all the faces in a particular set (and even faces outside this set) in a common space, that is, as weighted combinations of the basis vectors, or principal components. It also sustains the key proposition of the thesis, which is that the similarity of two faces is a function of the multidimensional distance separating them in 'face space'. The closer the faces are to each other, the higher their similarity. The representational basis also allows the definition of a related measure, namely facial distinctiveness. Facial distinctiveness is treated in Valentine's theory as the distance from the origin, or centroid, of the space, and this is easily operationalized in terms of the principal component basis. In sum, the PC (principal component) schema supports a similarity and distinctiveness metric in an effortless and natural manner, and is accordingly adopted for investigation in the empirical work of the thesis.

There are several aims in the empirical work, which is reported in Chapters 7 and 8. The central aim is to investigate the validity of the PC-based similarity metric. Even as a principal component space provides an elegant solution to the problem of finding a representatonal basis to sustain the similarity metric, this representational basis is thoroughly arbitrary: basis vectors are chosen by principal component algorithms on the basis of statistical criteria,[6] and they need not correspond in any way to human perceptions and judgements of facial similarity. The proposed measure clearly requires empirical investigation. A number of such investigations were conducted, and are reported in the relevant chapters.

[6] They are chosen to satisfy two central statistical, or mathematical, constraints: namely that the set of basis vectors is orthonormal, and basis vectors are in addition chosen so that they (sequentially) resolve maximum variance.

Seven studies were conducted in total. The studies involved approximately 700 subjects, and pre-study collections of facial images involved another 350 subjects.

Studies 1 and 2 are reported in Chapter 7. These studies were exploratory investigations of the similarity and distinctiveness measures. Study 1 piloted the principal component analysis of images, and a similarity ranking task. Study 2 extended the pilot work reported as Study 1, and examined the correspondence between the similarity measure and subject rankings of facial similarity. It also investigated the correspondence between the similarity measure and a 'face pairings' task. Both procedures produced results which indicated some correspondence between perceptions of similarity and the PC-based measure, but these were of a mixed nature in the case of the ranking task. Several other 'validity checks' were built into the design of Study 2. A cluster analysis of images on principal component coefficients produced clusters which appeared to group 'like' faces, and separate 'unlike' faces. A discriminant analysis of pre-defined sex, race, and age groups, using principal component coefficients as classificatory variables, proved capable of discriminating these groups with a high degree of accuracy. In addition, several other methodological issues raised by the image analysis technique were investigated in this study, and are reported. Investigations of the PC-based distinctiveness measure, on the other hand, produced results which uniformly disconfirmed a correspondence between the measure and subject perceptions of distinctiveness.

Studies 3 and 4 are reported in Chapter 8. Inspection of the method and results of Study 2 suggested that there were several methodological uncertainties in that study, and corrections for these were built into the design of Study 3. In the first place, a relatively small, homogenous set of face images was used in the earlier study. Study 3 corrected this by collecting a relatively large and heterogeneous set of face images. Corrections were also effected to the similarity ranking task of Study 2, particularly in an attempt to structure subject judgements. Large inter-subject variation in ratings of similarity were observed in Study 2, and instructions were devised in an attempt to counteract this variation. Results from 11 distinct rating sequences showed a strong correspondence between subject ratings and the similarity measure. Facial distinctiveness was investigated again in Study 3, but manipulations involving the PC-based operationalization proved difficult to interpret, and a lack of correspondence between this measure and subject ratings was observed again, corroborating results from Study 2.

Study 4 used the rating task of Study 3, and again examined the correspondence of subject ratings of similarity and the PC-based distance measure. In addition, Study 4 introduced an important manipulation. Subjects were shown frontal photographs of subjects, or profile photographs, or both frontal and profile photographs, and asked to complete the similarity rating task. This manipulation

made it possible to assess i) whether PC-based measures of similarity derived from frontal views of faces correspond to PC-based measures derived from profile views of the same faces; ii) whether subject ratings of profile views correspond to subject ratings of frontal views; and iii) whether the PC-based measure of similarity correlates with subject ratings of similarity, in both frontal and profile, and combined frontal and profile, views.

Results in the case of i) showed that the PC-based similarity measure derived from frontal images correlated reasonably highly with the PC-based measure derived from profile images. Much research in the face recognition literature has implicitly assumed that frontal images are sufficient representations of faces (see the discussion in Chapter 5 on this issue), thus confounding 'face recognition' and 'picture recognition', and it is important to show that the PC-based measure of facial similarity generalises beyond just one viewing perspective.

Results in the case of ii) showed that subject ratings of profile views corresponded reasonably well to ratings of frontal views, but correlations here were uniformly lower than correlations observed between frontal PC similarity scores and profile PC similarity scores. I suggest that the correlations between the sets of PC scores are therefore sufficiently high.

Results in the case of iii) showed again that facial similarity scores derived from a principal component analysis correlate highly, on average, with subject ratings of similarity. This was not uniformly the case in Study 4, however: a manipulation involving PC-based distinctiveness complicated the pattern of results.

In all rating tasks employed in Studies 1-4, subject judgements of similarity exhibited substantial inter-rater variation. There are many possible reasons for this; I argue in Chapter 7 that the most unfortunate explanation, in terms of the consequences for the central empirical project, is one which posits that subject perceptions of facial similarity are inherently unstable, that individual ratings of the same face will fluctuate over occasions. Accordingly, in Study 5 I investigated the stability of subject similarity ratings over time, using a standard test-retest design. Subjects completed a similarity rating task of the kind used in studies 3 and 4, and after a period of three weeks completed a second similarity rating task, which contained the same target, and approximately half of the faces used in the original array. Correlations assessing test-retest reliability were acceptably high, even under conditions which were not promotive of this.

The ambition underlying the central empirical project of the thesis is to find a direct measure of facial similarity which will serve at the same time as a measure of lineup fairness. Although Studies 1-5 appear to show, in sum, that the proposed measure corresponds to human ratings of

facial similarity, none of these studies provides any information about the usefulness of the measure in relation to identification parades.

Study 6 accordingly moved the focus of the empirical work to the question of parade fairness. Eighteen lineups of varying target-foil similarity and varying target distinctiveness were created as materials for a mock witness experiment. Subjects in this experiment were provided with descriptions of three suspects, and were able, in general, to identify the suspects with comparative ease. The identification rate was, however, affected by both manipulations. In particular, the identification rate (measured as a proportion) appeared to vary in a negative linear relation to the PC-based similarity score i.e. as facial similarity increased, the proportion of subjects correctly guessing the identity of the suspect decreased. The relation of the identification rate to distinctiveness was not clear, once again showing that the PC based distinctiveness measure does not seem to be of much use - at least in its present form. Correlations between the measure of facial similarity and indices of lineup fairness commonly used in the literature (i.e. 'functional size', 'effective size', and the measure 'E', which is proposed in Chapter 6) were all strong, and in the expected direction.

In Study 7, the final empirical study reported in the thesis, I investigated the relation between facial similarity and identification performance, in a simulated identification scenario. This relation has rarely been examined in the literature: it is a central element in legal and psychological thinking about identification parades, and some empirical evidence about its nature may be useful. A randomised factorial experiment was devised, incorporating manipulations of facial similarity (operationalised in terms of the PC-based measure), perpetrator presence, and lineup structure (i.e. simultaneous vs. sequential). Subjects were asked to examine biographical information, and photographs, of three fictitious students at their university, and after a period of half an hour were tested with photo-lineups, which were constructed to embody the manipulations described above.

Results showed several things. In the first place, the superiority of sequential lineups over simultaneous lineups - well established in the literature - was confirmed, and in the expected manner. That is, sequential lineups led to fewer mistaken identification decisions when the perpetrator was absent, but to no fewer correct identifications when the perpetrator was present. Identification accuracy was strongly affected by target-foil similarity, but in an unanticipated fashion: lineups which had the lowest degree of target-foil similarity produced the highest degree of correct identification decisions, both in situations where the perpetrator was present, and in situations where the perpetrator was absent.

The relation between similarity and identification accuracy is not inexplicable, taken on its own: in Chapter 4, I discuss the proposition by Wells and colleagues (Wells, Rydell & Seelau, 1993; Wells, Seelau, Rydell & Luus, 1994) that lineups should be constructed to have 'propitious heterogeneity', i.e. target-foil similarity should not be too high. There is a useful *reductio ad absurdum* here, which is to imagine a lineup constituted by a suspect and nine of his clones: such a lineup will yield a highly inaccurate identification rate, and this implies some optimal suspect-foil similarity function.[7]

However, when the results of Study 7 are combined with the results of Study 6, there appears to be an interpretative difficulty of some magnitude. Study 6 shows that an increasing degree of facial similarity leads to fairer lineups (as determined in mock witness tasks), but Study 7 shows that greater similarity leads to fewer correct identification decisions. This implies, counter-intuitively, that fairer lineups lead to poorer identification accuracy.

I will return to this problem in the final chapter of the thesis, to wit Chapter 9. The rest of the thesis sets the stage for this return.

[7] Wells and colleagues, however, do not think that it is useful to pursue the notion of an 'optimal similarity function'. See the discussion in Chapter 4.

Chapter 2 Applying Psychology

Introduction

This thesis examines ways of measuring the fairness of police identification parades, and is therefore an exercise in applied psychology. Methodological and measurement issues are also of some importance, but are subsidiary in most respects to the central concern with identification parades, and the testimony of eyewitnesses.

I will argue in this chapter that research on eyewitness testimony depends for its justification on the prospect of application. The research derives from a concern with a practical problem, namely the consequences that honest, but mistaken identifications have for the fair discharge of justice. Pre-trial identifications are characteristically secured with the assistance of identification parades, since it is widely believed that they make such identifications more reliable. Documented cases of mistaken identification have shown that they do not eliminate the problem (Loftus, 1979; Loftus & Ketcham, 1991), and it is perhaps not even clear that they significantly alleviate it. Psychologists have addressed research in the last twenty years to ways in which these 'recognition tests' can be improved.

This interest in eyewitness identifications stems from a more widespread, and recent, concern with applying psychology to 'everyday' problems. The rush to do 'applied psychology' is often accompanied by an unfortunate neglect of the conceptual issues that underlie the notion of applied research. I will consider these in some detail here, and take a few observations in this respect as rough guides for later chapters. I begin by considering reasons for the contemporary enthusiasm for applying psychological research outside of laboratory settings.

'Crises' and the drive to application

As long as most psychologists can remember, psychology has been in a state of crisis.[1] Each generation re-asserts the existence of the crisis - usually as if the crisis had just been discovered - and

[1] Stehr (1992) argues that the pronouncement of 'crisis' recurs in almost all of the social sciences, from the early part of the twentieth century onwards. Generally, these crises devolve on the (cyclical) assertion that social science knowledge of the period has failed to make practitioners competent to understand or solve contemporary social problems.

re-defines its nature, with an eye to justifying its own solution to this intractable state. There is remarkable agreement about the existence of the crisis, especially since psychologists disagree about so many other things!

One statement of the crisis, which we associate with Tajfel, Israel, Gergen and others, from the late 1960's, objects to the narrow conceptualization of both method and theory that mainstream social psychology exhibits. Laboratory experiments drive the discipline, and these are conducted as if in a vacuum, hence the title of Tajfel's (1972) paper 'experiments in a vacuum.' This cannot be good for the discipline; the arguments are well known enough to pass them over here. Of course, more radical statements of the crisis can be found, of almost any description. Séve's Marxist statement of the crisis is an example (Séve, 1976), as is the work by the 'Loughborough group.' The title of Ian Parker's recent book (1989) is indication enough of the prevalence of the view that the discipline is permanently besieged: *The crisis in modern social psychology and how to end it.*

I do not want to labour progress here with an examination of the many formulations of the alleged crisis. I will look at one in detail. This is the view that psychology fails to concern itself with real problems, indeed with everyday life. There are many variants of this view; I would like to present the position adopted by the American cognitive psychologist Ulric Neisser, because his statement of crisis has led directly - and indirectly - to a flourish of research that seeks to concern itself with everyday life.

The genesis of applied cognitive psychology - Neisser's polemic

Almost two decades ago, Ulric Neisser (1976), a leading cognitive psychologist, and author of a seminal text that spearheaded the 'cognitive revolution' (Neisser, 1967), seriously questioned the scientific validity of cognitive research. He claimed that cognitive psychology lacked 'ecological validity', that it failed to acquire relevance outside of the laboratories in which it was constructed. Where psychoanalysis and behaviourism insured themselves, during their periods as hegemonic discipline, by making themselves applicable, cognitive psychology had failed to secure itself in this way. Neisser recommended a radical change in cognitive psychology's orientation if it wanted to endure as a psychological discipline - he predicted a rapid demise if it failed to apply itself outside of its laboratories.

> Every age has its own conceptions - men are free or determined, rational or irrational: they can discover the truth or they are doomed to illusion. In the long run, psychology must treat these issues or be found wanting. A seminal psychological theory can change the beliefs of a whole society as psychoanalysis, for example, has surely done. This can only happen, however, if the theory has something to say about what people do in real, culturally significant situations. What it says must not be trivial, and it must make some kind of sense to the participants in those situations themselves. If a theory lacks these qualities - if it does not have what is

> nowadays called ecological validity - it will be abandoned sooner or later. (Neisser, 1976, p 2).

Several years later, Neisser (1983) pursued this argument with an edited collection of memory research papers that addressed problems in ways that were clearly more applicable to conditions outside of the laboratory. He chose the papers as examples of what directions cognitive psychology should follow if it wanted to do 'ecologically valid' research, and commenced the collection of papers with a restatement of his 1976 argument, this time with the psychological study of memory particularly in mind.

> I think that memory in general does not exist. It is a concept left over from a medieval psychology that partitioned the mind into independent faculties thought and will and emotion and many others with memory among them. Let's give it up and begin to ask our questions in different ways - our questions need not be uninformed by theory or by a vision of human nature but perhaps they can be more closely driven by the characteristics of ordinary human experience (Neisser, 1983, p 12).

The task is thus to make the study of memory applicable to 'ordinary' lives. That knowledge generated by cognitive psychology might be used for practical gain is a theme that recurs in *Memory Observed*.[9] One section of the collection, for instance, is devoted to cognitive psychological research on eyewitness testimony, and Neisser himself contributes one of the papers to this section of the book. Here the point is stressed that research can very usefully be redirected for obvious and important practical gain.

Neisser dismisses the 'laboratory' approach to the study of memory on the grounds that in the hundred years of its existence it has accumulated very little knowledge about the experiences we ordinarily think of as involving memory processes. Its findings are restricted not only in terms of generalizability - scenarios typically used to study memory in laboratories rarely resemble those outside the laboratory - but their theoretical representation is usually also restrictively experimental. (Thus, 'memory interference' *means* performing in a particular way on a list learning task). Laboratory research on memory has failed to accumulate the kind of findings that constitute a coherent and useful body of knowledge. What knowledge it has accumulated, anyway, presents little advance on what is immediately obvious - indeed, on what is immediately obvious to pre-schoolers.[10]

The strongest and most notorious assertion by Neisser, however, was the sheer irrelevance of memory research:

---

[9] Neisser was not the only cognitive psychologist at that time who pursued this line of critical inquiry, but his arguments appear to have had the greatest influence. Alan Baddeley, at the M.R.C. unit in Cambridge, made a similar argument, somewhat more whimsically, as is apparent from the title of his 1988 article (Baddeley, 1988).

[10] This is not mere sarcasm on Neisser's part. Several studies have shown convincingly that children of pre-school age know the important experimental findings in traditional psychological laboratory research. (Kreutzer, Leonard and Flavell, 1975).

> If X is an interesting or socially significant aspect of memory, then psychologists have hardly ever studied X (1978, p 4).

The solution lies in changing the content of memory investigations. Neisser urges in several places that memory research address 'practical problems':

> ...some of the best minds in psychology have worked and are presently working in the area of memory. Why then have they not turned their attention to practical problems and natural settings? (1983, p 6)

The message is clear in *Memory Observed*: cognitive psychology lacks ecological validity and one of the ways of doing 'ecologically valid' cognitive research is to do research that can be used outside of the laboratories in which it is produced. It is to do what is commonly called 'applied psychology.'[11]

Neisser's argument has been very influential, perhaps especially in the area of immediate concern to this thesis, eyewitness testimony research. Neisser specifically singled eyewitness research out for praise in his 1978 paper: he considered it an area of cognitive psychology that concerned itself with a real problem. It is also an historically interesting tradition of research as far as applied psychology is concerned, since the earliest applied psychologists took it as one of the obvious areas for an applied psychological science. Hugo Münsterberg, often called the father of applied psychology, wrote a book and several papers on the subject in the first decade of the twentieth century (Münsterberg, 1908). Stern (1910), Whipple (1912) and many others, in continental Europe, in England, and the U.S.A., did likewise.

This early research petered out in the second and third decades of the century, and there was very little recently published work at the time of Neisser's statement of the crisis. After Neisser's message, the area boomed. Between 1979 and 1986, at least seven books on the subject were published (see, for examples, Loftus, 1979; Yarmey 1979; Wells & Loftus, 1984; Lloyd-Bostock and Clifford, 1983), and by 1994 this number had risen to close to twenty five. Over 1000 journal articles appeared between the years 1977 and 1995.[12] There are presumably many reasons for this explosion of research, but I think Neisser's message to 'go forth into the world' resonated loudly down the corridors of academic psychology, and many research psychologists rallied to the call.

However, Neisser's polemic has not gone unanswered. It is worth digressing briefly to consider a recent counter by the Yale psychologists, Banaji & Crowder (1989). The value of this exercise derives from the fact that the chief interest of the research presented in this thesis is the 'real-world'

---

[11] Of course, Neisser's argument is much lengthier and more sophisticated than the brief synopsis presented here. I have adumbrated and simplified it in order to emphasise the concerns with ecological validity and application.

[12] A more detailed argument is technically required here: I offer these quantitative estimates on the basis of personal knowledge of the literature, and admit the possibility that they may be slightly inaccurate.

witness problem of identification parades, and there are important conceptual questions hidden in the notion of applying psychology to such problems.

The Response to Neisser: Banaji & Crowder (1989)

There are two major propositions in the argument made by Banaji & Crowder (1989). In the first instance, they assert that the move out of the laboratory to the locus of naturally occurring memory phenomena is misplaced. All sciences study phenomena in laboratories, and then attempt to apply acquired knowledge to phenomena in their naturally occurring state. The notion that a scientific enterprise can make its results more ecologically valid by studying phenomena in their natural form is without sound precedent: it is like imagining an astronomy conducted with the naked eye (Banaji & Crowder, p 1185). Since so many other scientific disciplines manage to bridge the interpretative gap between the laboratory and the natural world, there is no *a priori* reason why memory research should not be able to do so.

In the second place, there is no reason why a mere concern with 'everyday phenomena' should render a piece of research ecologically valid. In this respect, Banaji and Crowder list many studies that include samples or tasks because of their interest value, or their 'eccentricity', and for no other scientifically respectable reason. The point is that these studies survive on an illusion of ecological validity. They aim at a high ecological validity of method, but succumb to low generalizability of results. The multiplicity of uncontrolled factors in naturalistic observation prohibits generalizability to other situations; tests in the real world do not permit generalizability since variability in real world situations is immense.

It is better, argue Banaji & Crowder in conclusion, to have results that are generalizable, even if the ecological limitations on this generalizability are unclear. Results that are not at all generalizable, despite a superficial veneer of ecological realism, are of no use at all.

Replies to Banaji & Crowder

The 1989 article by Banaji & Crowder appeared in the mouthpiece of the American Psychological Association, *American Psychologist*, and generated heated responses.[13] These were assembled in a

[13] The arguments made by Banaji & Crowder are not new, and neither are the responses. A similar dispute, with similar arguments for and against, appeared in the American Psychologist in the early 1980's: see, for example, Berkowitz & Donnerstein's (1982) 'answers to criticisms of laboratory experiments'. The cyclical occurrence of such disputes is in line with the remark I made earlier about the cycles of 'calls to crisis' which dominate social scientific literature and research.

discussion forum in the 1991 volume of the same journal, and it is worth repeating some of the key points.

It is not clear that the analogy with the physical sciences drawn by Banaji and Crowder supports their contention: although chemistry may be a particularly successful laboratory science, there are many other sciences (to wit, Zoology) for which field research is invaluable, and for whom its absence would spell severe retardation (Neisser, 1991). Furthermore, there are many memory phenomena which are clearly inscrutable by laboratory investigation: Bahrick's (Bahrick, Bahrick & Wittlinger, 1975; Bahrick, 1984) 50 year follow up studies for memory of material learned at school is a case in point.

Furthermore, it is a mistake to characterize the difference between laboratory research and the research of the so-called 'everyday memory school' as one of method (Klatzky, 1991). Both traditions of enquiry share a belief in many common methodological canons, but are effective in different situations.[14] It is just a mistake to ignore the fact that laboratory guided memory research has very little to say about ordinary memory phenomena (Morton, 1991), and appeals to 'guaranteed' methods of attaining generalizability are sorely tested by the patent lack of success in over 100 years of laboratory research.

I am reluctant to argue here for a settlement of the dispute: it is a long-lived and remorseless tyrant of a problem, and has probably marked out its permanent place in the psychological gallery. What I wish to assert instead is that there are problems other than those raised by Banaji & Crowder with the endeavour to take memory research out of the laboratory. A number of very thorny issues lie below the surface prospect of 'solving practical problems' with research. These need to be considered closely if we are to take seriously the suggestion that 'application' is a way of securing 'ecological validity' for a piece of research.

As the research to be examined in this thesis takes much of its impetus from the prospect of applying its findings to legal matters, I will deal with the issues that 'application' raises at some length.

The notion of an 'applied psychology'

'Applied psychology' is one of the oldest formally recognized and institutionalized branches of psychology, at least in name. As early as 1908 a chair in Applied Psychology existed in an American institution (the Carnegie Institute of Technology), and by 1917 the American Psychological

[14] Davies (1992) has more recently argued in favour of methodological pluralism, particularly with respect to research on witnesses.

Association had established the Journal of Applied Psychology. Thus, by the early decades of this century, applied psychology was considered to be a firmly established subdiscipline: institutes existed for applied psychological research, chairs of applied psychology existed in several universities, and a burgeoning technology (mental testing) was associated with applied psychology.[15]

For the purposes of the discussion here, some idea of what 'applying psychology' means, at least in conventional terms, is needed, and so I will introduce a number of conceptualizations from prominent 'applied psychologists'.

Dudycha (1963) likens the applied psychologist to an engineer: in the same way that the engineer applies theoretical physics, the applied psychologist applies the findings of psychological science. This allows the rather mundane reading of Applied Psychology as 'psychology in use':

> . . . the application of psychology to the various areas and aspects of individual and social life. p 4)

Notice how Dudycha displaces the onus of the justification: applied psychology is sound insofar as psychological science is sound.

This link between application and the extension of the scientific method is made even clearer by Anastasi (1964). The applied psychologist's contribution stems from 'his research approach to the problems of human behaviour': he takes the scientific method common to all sciences into applied contexts. Insofar as applied and basic research differ, the essential differences concern:

1. How the problem is chosen. Basic research chooses problems to help in the construction of theories, applied research chooses problems to help in administrative decisions.
2. The specificity and generality of results. Basic research is more generalizable; in applied research generalization is of limited validity (as applied research is less concerned with theoretical and causal relations). (p 5).

The problem, for Anastasi, is to distinguish 'applied' from 'pure' without sacrificing the legitimizing claim on the 'scientific method'.

Both conceptualizations considered thus far attempt to distinguish applied science from pure science formally: the difference between the two endeavours resides finally in the locale of practice, and whatever other differences there are may be specified from knowledge of this fact. In much the same vein is the following conceptualization (van der Vlist, 1982), which appears in an influential European series of monographs.

---

[15] The historical account presented here is taken from a reading of the accounts given by Anastasi (1964) and Dudycha (1963). As an historical presentation it is quite obviously inadequate, but it does make the point fairly clearly that Applied Psychology has long been considered a respectable subdiscipline of psychology.

According to van der Vlist, applied and pure research differ in the way they treat their respective independent and dependent variables. Where pure research concerns itself principally with the relationships between independent and dependent variables, applied research is seen to be concerned exclusively with a particular dependent or independent variable. Pure science serves to increase our knowledge about a theoretically interesting relationship; applied science serves as the basis for decision making with respect to a concrete dependent or independent variable.

The relationship of this view to the views espoused by Anastasi & Dudycha is fairy transparent and the observations I made earlier apply equally here. I single out van der Vlist's treatment because it makes the pointed claim that 'applied research' is so named because of its application: later I will suggest that the 'naming' of applied research is a much more arbitrary matter than this.

Applied research is the extension of pure research into 'ecologically real' locations. This is the line of argument pursued in all of the accounts I have considered thus far, but subtler formulations are naturally possible. So, for example, Warr (1978) denies the applied/pure dichotomy, and instead asserts that it exists as a dimension, stretching from the completely pure investigation to the completely applied project. Where, on this dimension, a particular psychological enterprise falls, depends on (i) the population studied, (ii) the research setting, and (iii) the intended outcome of the work. The problems to be studied are taken directly from real life situations, although they may be studied on the spot or in laboratories.

> the applied nature of an investigation derives from the fact that its proximal origin is in a general sense external to the discipline (p 11)

whereas

> Pure psychological research aims to deal with an issue raised by the results, theories or ideas of psychologists themselves, being part of a short cycle feedback system feeding directly upon its own outputs (p 11).

I have considered a number of attempts to identify the intellectual pursuit that constitutes 'applied psychology'. I want to show in the next section of the chapter that these accounts take us no further in our attempt to understand the nature of applied research, and I will suggest there that the notion of an applied psychology is consequently better treated as a meta-theoretical problem than as a category of research.

Problematizing conceptions of applied psychology

In a provocative and incisive paper, Jonathan Potter (1982) argues that most discussions of applied science can be subsumed under a more general ideological practice which attempts to present 'science' as socially useful, as the origin of many of the things that improve our lives. The notion of

'applied science' serves this broader function by contributing to it an 'ideology of application' (1982, p 24): the intimate relation held in scientific cultures to exist between science and technology. The first two conceptualizations of applied psychology that I discussed - those espoused by Anastasi and Dudycha - are exemplary instances of this: applied psychology is the transportation of the scientific method to locations proximal to the discipline.

This conceptualization does not stem from the psychological literature in particular. Indeed, it is a prevalent way of thinking about research in most scientific disciplines, and is known in the philosophy of science as the 'two stage model' of scientific activity (see Danziger, 1990, who traces the rise of its popularity in the late nineteenth and early twentieth centuries). This is the familiar distinction between 'basic' and 'applied' science. The task of basic science is the painstaking construction of universal theory, and applied science sees to it that this universal theory is applied outside the laboratory. The conceptualization specifically indicates a direction of information flow, namely from basic to applied.

But just how close is the relationship between science and technology? The suggestion of an intimate relation between science and technology - as inscribed in the ideology of application - is, to say the least, problematic: at any rate, the relationship is not of the direct form suggested. An increasing body of research in the sociology of science and philosophy of technology suggests that technology is not simply applied science (in the sense given to the term in the ideology of application). For example, research on the USA weapons industry shows that 91% of innovations in the technology originated from inside the technology itself, and only 9% from scientific research (Potter, 1982). Similarly, studies using citation analysis find that

> Science seems to accumulate mainly on the basis of past science, and technology primarily on the basis of past technology. (Mulkay, quoted in Potter, 1982, p. 23)

This is not to say that technology and science bear no relationship to each other: the sense in which technology and science do relate is best taken as a case of enablement, but this enablement is in a direction contrary to that hypothesized by the ideology of application. Ihde (1979), for instance, argues that the history of technology shows that technology of a particular form is a prerequisite for science of a particular form.[16] Thus, watermills (among other technological innovations) existed before, and were prerequisites for Newtonian mechanics. That knowledge "flows" from the pure pole of the pure - applied dimension to the applied pole is an untenable thesis. There is a case to be made for connections between basic and applied science in particular instances, but there is certainly nothing of the suggested dependency.

[16] For a similar argument in respect of mental testing and its relation to theories of intelligence, see Danziger (1990).

The claim serves clear ideological interests: connotations about the social utility of 'science' slip into the way we think about our lives. It is probably to maintain the implications attendant upon the idea of a flow of knowledge from 'pure' to 'applied', that prevailing conceptions of 'applied psychology' identify the origin of the research problem as the feature that distinguishes pure from applied research.

The matter does not end here. Potter makes an important distinction between applied psychology and applicable psychology. The point is that most of what we call applied psychology is really only applicable psychology: findings made under the name of applied psychology are generally not applied - the label is assumed only because of a superfluous concern with issues or problems in society. In most 'applied research', all that happens is that researchers pluck problems from the outside world and justify their work in academic terms. The research is called 'applied research' only because it (ostensibly) addresses a social problem.

This discussion of the problems in the orthodox conceptualizations of applied psychology has drawn attention to the ideological notion of the pure-applied split and the pertinent distinction between applied and applicable psychology. There is a further point I wish to make about these conceptualizations, which is the way in which the 'applied' in 'applied psychology' is treated.

The 'applied' in applied psychology, is read, in all the accounts I have presented here, as the name of a subdiscipline of Psychology. To put it clearly: the question of an applied psychology is treated in terms of what makes applied psychology a discipline, not in terms of why applied psychology is an applied endeavour. Instead of focusing on how psychology is applied, elaborate attempts have been made to show how applied differs from pure. This is a type of nominalist error: because a name exists, an entity is assumed to correspond to it. In this way psychologists have taken for granted the 'applied' nature of applied psychology, and have failed to ask important questions about *if* psychology is applied, and *how* it is applied. Thus what is really only applicable psychology at best has come to constitute what goes by the name of applied psychological research.

So we can see that the notion of an 'applied psychology' is a very problematic one indeed. Neisser's recommendation that memory research apply itself to practical problems then, is certainly not the guarantor of a new scientific respectability that he suggested it might be.

This does not mean that applying psychological research, or doing applied psychological research, is impossible. Examples of situations where research findings have been applied, and situations where they have been extremely useful, are not difficult to find. Thus, the recent legal revision of rules covering child witnesses in the U.K. used a review of the psychological literature as a preface (Davies, 1992). Similarly, the collective of researchers informally known as 'the Aberdeen group'

has produced an artificial facial composite production system that is currently in use in several U.K. police forces (Ellis, J.W. Shepherd, J. Shepherd, Flin & Davies, 1989).

It does mean, however, that we need to think clearly about the nature of the research that we engage in when we study 'problems' of the sort recommended by Neisser. The choice of problem often serves as justification for the research, and it is consequently very important to have a clear understanding of the nature of the problem in its original context, as well as some idea about the concrete *application* of the ensuing research. Eyewitness researchers have frequently failed to consider these specific issues, but there are several exceptions. In the next section, I will examine a few of these in the hope that they may inform the approach to be adopted in this dissertation.

Types of 'eyewitness variable'

Much eyewitness research is conducted with little attention to the generalizability of findings across situations (Read & Bruce, 1984). There are many ways in which one can become an eyewitness to an event, and it is not clear that different witness-creating events have enough in common to justify treating them with any degree of equivalence (Bekerian, 1993).

It is accordingly important to differentiate witness scenarios according to the variables of interest, and to deal with them separately. A simple and very useful distinction in this regard was made by Wells (1978). Wells argued for a crude dichotomization of variables at play in the identification process. *Estimator variables* are attributes of the witness or the event which cannot be controlled. The physical lighting of the crime scenario, the length of time the witness observed the perpetrator, and the gender of the witness, are examples. These variables may all affect the accuracy of the testimony, but since they cannot be controlled, one can only estimate the effect that the variables have on the accuracy of the witness's report. Such an estimation is of course an extremely imprecise matter, and those who opt to study them need be mindful of this. *System variables* are attributes of the legal system - and perhaps of the event, or witness - which affect the accuracy of the identification, and which can be controlled and modulated. The type and composition of an identification parade, and the style of interrogation used by a police officer are examples of such attributes. The idea is that we can improve the quality of witness testimony by identifying and researching system variables. Table 2.1 presents a simplified classification of aspects of the identification process according to the distinction between the two types of variable.

Estimator variables		System variables
Situational factors	*Witness factors*	*Parade factors*
Physical conditions	Race	Composition
Type of crime	Sex	Functional size
Stress and arousal	Age	Instructions
Duration of incident	Confidence	Parallel parades
Delay effects	Intelligence	Social dynamics
Interpolated activity	Personality	Mode of presentation
Extraneous information		

Reproduced from Shepherd, Ellis & Davies, 1982, p 7.

Table 2.1 Estimator and system variable classification of eyewitness identification factors

The underlying implication is that the application of research should be facilitated by studying system variables. This is important to Wells, who asserts

> ... in undertaking an applied project it is incumbent on a researcher to demonstrate the applied utility of an eyewitness study. (Wells, 1978, p 1555).

The distinction between estimator and system variables is astute, and is often used in review discussions of eyewitness research. It is taken up by Deffenbacher (1991), for instance, and by Shepherd, Ellis & Davies (1982), who use it to justify the classification presented above as Table 2.1. It is an astute distinction because it achieves classification of a myriad number of variables, while emphasising the point that eyewitness research depends for its justification on the prospect of application. The object is *not* to develop a comprehensive theory of the eyewitness; indeed such a project cannot hope to succeed.

Nevertheless, studying variables that facilitate application is only a step in the right direction. The gap between research knowledge and application of that knowledge is daunting, even if the research is tailored for the prospect of application. In the next section we consider specific methods of application aimed at bridging the gap.

Models for applying eyewitness research

Eyewitness research has entered the legal system in several forms, and with varying success. Again, Gary Wells has made a pertinent distinction (Wells, 1986), this time in order to distinguish the ways in which eyewitness researchers can hope to secure application of research knowledge.

Expert testimony

The first and perhaps most controversial route of application is by expert testimony in court trials. The idea here is that expert testimony can help 'triers of fact'[17] identify relevant variables and understand their influence on eyewitness testimony. Presumably, too much or too little credibility is ascribed to eyewitness testimony in court trials, and expert testimony helps to correct this. Experts could testify with respect to estimator variables, in which case the testimony is aimed at providing a sort of 'social framework' for the triers of fact, or the expert could testify about system variables, and this may include giving an opinion or measure of the fairness of identification procedures used in the particular case.[18]

Expert testimony is not new to the courts, and there are articulated criteria governing the admittance of such testimony. However, Monahan & Walker (1988) argue that these criteria are neither coherent nor consistently applied,[19] and the experience of several researchers in the field is that the admittance of expert testimony on eyewitness issues is a hit and miss affair (Loftus & Ketcham, 1991).

There are also other, fundamental problems with expert testimony on eyewitness research. These have been the basis for heated debate over the last 10 years (McCloskey & Egeth, 1983; Loftus, 1983a, 1983b; Egeth, 1993), and will be mentioned briefly here.

Expert testimony seems to reduce the tendency of jurors to believe eyewitnesses (Wells, 1986). However, if jurors are not prone to 'overbelieve' witnesses in the first place, this testimony may be inappropriate. Furthermore, even if it appropriately reduces such 'overbelief', it is not clear whether such testimony makes jurors more accurate at distinguishing correct and incorrect witnesses: a simple, blanket correction of the 'overbelief' may not make jurors any more accurate, but may merely make them indiscriminately skeptical (McCloskey & Egeth, 1983). It is also not clear that estimator variable research will tell jurors anything they do not know: McCloskey & Egeth (1983)

---

[17] In the U.S.A., this function is usually served by a jury. In South Africa, there are no jury trials, and the presiding judicial officer(s) are the triers of fact. A distinction is still made, nevertheless, between matters of fact, and matters of law.

[18] There are ways of having testimony admitted to trials other than the traditional *viva voce* route. The so-called 'Brandeis brief', or *Amicus* brief, is a method suggested by Melton (1987), as a model for entering psychological knowledge into legal trials. The idea is that a professional body (such as the APA) acts as a 'friend of the court', submitting written argumentation based on scientific knowledge.

[19] Monahan and Walker make a case for a new model of treating social science research in courts, which they call the 'social framework' approach. The distinction between legislative and adjudicative fact, currently used in deciding on the relevance and admissibility of social science research, is jettisoned in favour of 'social authority', the central concept in the new paradigm: i.e. social science research should be treated as a source of authority rather than of fact. Social science research should be used just as precedent works in the common law: that is, it is binding on lower courts, the coherence of the argument is the criterion for evaluating it, acceptance and application by other courts is a recommendation for use, etc. Judges could do the research themselves, and psychologists could present evidence by deposition and not necessarily *viva voce*. The justification offered has various elements, but is based in particular on the claim of an overarching similarity of social science research and law: both are general, and produce principles, but there is also much deliberation required for particular cases. Monahan and Walker point out that several courts in the USA have started using social science findings as *social framework* evidence: i.e. the use of general conclusions from social science research in determining factual issues in a specific case.

suggest that much estimator variable research simply reinforces widely held beliefs about eyewitness performance.[20] Finally, how are jurors to use the information delivered in the testimony? Each juror will need to internalize the testimony, match the facts of the case against known research, and then estimate the effect that various aspects of the case are likely to have had on the accuracy of the witnesses. In all likelihood, though, it will do little more than make them consider the eyewitness identification somewhat more carefully.

Other criticisms of expert testimony on eyewitnesses worth mentioning here include i) questions about the quality of estimator variable research, and ii) positions which dispute the need for the research in the first place. Thus, McCloskey & Egeth (1983) argue that research on many estimator variables is contradictory and inconclusive, and that what is known about such factors is usually restricted to first and second order effects - almost nothing is known about higher order interactions. Konecni and Ebbesen (1986), on the other hand, argue that eyewitness identifications are admitted as evidence in such a small percentage of criminal cases that the problem is not worth bothering about.

These criticisms are themselves not inviolate, and several detailed a priori and empirical studies have attempted to demonstrate their flaws. Thus, Goldstein, Chance & Schneller (1989) surveyed judicial districts in the U.S. and estimated that eyewitness identifications are admitted in about 77 000 trials per year; Kassin, Ellsworth & Smith (1989) surveyed eyewitness experts and found a substantial level of agreement on most research findings; and Elizabeth Loftus has replied on several occasions to many of the other criticisms (Loftus, 1983a, 1983b, Loftus & Ketcham, 1991).

This is not the place to attempt an arbitration of the dispute. What I wish to take to the next section of the chapter is the question of the use to which jurors are expected to put expert testimony. No-one has contested that this requires a very difficult act of estimation on the part of the witness, and it may have a major constraining effect on the efficacy of eyewitness research.

System variable research

The second category Wells (1986) considers, in his classification of ways to apply eyewitness research, is a set of three strategies aimed at securing application for system variable research.

[20] For example, that the opportunity and length of observation are important determinants of accuracy (Deffenbacher, 1991); that lengthier passages of time between observation and identification lead to more mistakes (Wells, 1993), and so on.

Publish effectively

Since the ordinary business of researchers is to publish their findings, a useful strategy may be to publish in journals and books that are likely to be read by those who control the system variables under investigation. Melton (1987), for example, provides compelling evidence that most extra-precedent material cited in legal cases is originally published in law journals. There is very little use of non-legal material. The model of application here is a slow process of diffusion of psychological knowledge into legal practice. However, as both Wells and Melton concede, editors of law journals are not usually receptive to experimental research, and it is not clear what proportion of the 'legally enabled' read these journals. There is no guarantee of application by this route.

Incorporate knowledge of system variables in police training

The second strategy identified by Wells is the incorporation of knowledge into training at police training centres. Since police officials are frequently at the first site of work involving eyewitness identifications, this could be a particularly potent method of application. Unfortunately, it may be an impracticable strategy. Police services do not presently place much importance in training officers in identification procedures, and indeed usually offer no training in lineup administration. This is assumed in many services in the U.S.A. to be self-evident (Wells, 1986), and much the same appears to be true in South Africa.[21]

Change formal procedures and rules of conduct

The third strategy suggested by Wells is to aim at changing formal procedures and rules of conduct. There are several possibilities here. Statutory regulation of identification procedures could be enacted at the level of law, and research knowledge could inform such a process. Judicial commissions have considered the problem of identification evidence in several countries this century,[22] and although psychological research findings have often been considered, they have had little force. In the most recent English commission, chaired by Lord Devlin, the committee expressed the opinion that psychological research was not presently of much assistance, but nevertheless, that

> ...the possibility should be explored of undertaking research directed to establishing ways in which the insights of psychology could be brought to bear on the conduct of identification parades. (Devlin, 1976, cited in Shepherd et al., 1982, p. xi.)

---

[21] This was communicated to me and to a legal colleague (Shereen Volks), in a telephonic interview we conducted with South African police officers in Cape Town in 1991.

[22] In England, no fewer than three judicial commisions have investigated the problem, and made recommendations (Devlin, 1976). In Canada, at least one commission has been briefed with a similar task, and has submitted a report (Brooks, 1983).

Alternatively, psychologists could aim at securing strong procedural recommendations from legal bodies. Research on identification procedures would form part of the basis for these recommendations.

Again, Wells is not optimistic about the prospects of successful application. Statutory regulations or procedural recommendations are unlikely to be promulgated, since police practice is traditionally treated in Anglo-Western traditions as largely independent. Whatever checks are necessary will emerge in the evolution of the common law through judicial precedent. In South Africa, for example, several judges have stressed that the conduct of identification parades is a matter of police practice, and the power to conduct identification parades is enacted at statutory level (see Chapter 3 of this thesis).

Problems with the system variable approach to application

Although each of the strategies for applying system variable research has some promise, closer inspection shows that there are obstacles in the way of each. Wells (1986) is pessimistic about the prospect of application by these routes, and suggests that a better overall strategy may be to provide expert testimony on system variable research in particular legal trials. I will examine his argument in favour of this claim in the next section of the chapter, but wish for the moment to dwell briefly on the discredited strategies

With the benefit of hindsight, it appears that Wells' pessimism may have been misplaced. There is evidence now to suggest successful application of some of the major system variable research. Two examples will suffice. One of the most robust and interesting findings in research on identification parades is the plain superiority of the sequential identification parade over the traditional simultaneous identification parade (see Chapter 4 for technical details). This form of identification parade was apparently developed by Rod Lindsay, in collaboration with Gary Wells (Lindsay & Wells, 1985), and leads to fewer false identifications, with no significant decrement in the number of accurate identifications. North American police forces have taken note of the research and it is now estimated that some 15% of identification parades are of the sequential form (Malpass, personal communication). Clearly, successful application has occurred here, and it appears that effective publication of the results has been one of the proximal determinants of the application.[23] A second example of successful application is the development of a computerized facial composite system by the 'Aberdeen group'. Prior research pointed to the inadequacy of the traditional police methods for

[23] It is difficult, of course, to provide firm evidence for this proposition, and should perhaps be considered a tentative hypothesis about the determinants of the application.

producing facial composites (i.e. the commercially produced systems Identikit, PhotoFit, and sketches by police artists). In one study, for example, it was shown that police officers charged with composite production were not able to produce a convincing composite of a well known face, even with a photograph of the person in front of them (Ellis, 1984). The 'Aberdeen group' has in the interim period produced a computerised system that produces composites of palpably higher quality, and which also appears to lead to better accuracy rates in experimental settings than the traditional techniques. This system is now in use in several British police services, and police are given special training in composite production. Psychological research on identification issues has clearly achieved successful application here, and this has occurred at the level of change in police training.

Nevertheless, there is much system variable research that has not met with application - perhaps *most* has not met with application - and it may serve us better to understand the reasons for failure to achieve application, than to derive faith from specific instances of successful application. This is a larger project than I am able to undertake in this dissertation, and I will settle instead for the few rough pointers in the right direction, outlined in the final section of the chapter.

Expert testimony on system variables

Wells (1986) concluded at the time of writing that neither of the two broad approaches to application of witness research had much success (expert testimony on estimator variables, the three strategies aimed at application of system variable research). He argued that it might be more effective to combine the approaches. In particular, he suggested that testimony on system variable research could be very effective, but doubted that the same could be true of testimony on estimator variable research. This is because testimony of the latter sort could only be descriptive of broader issues of eyewitness reliability, whereas testimony of the former sort effectively makes proscriptions on police practice. Testimony on system variable problems is likely to have a chain effect - if an identification is procedurally discredited by testimony, prosecutors will lean on the police to change their practice, and this will result in modifications still further down the hierarchy. Prosecutors are after convictions, and there is thus a built-in incentive to change. In the case of testimony on estimator variables, however, there *is* nothing either prosecutors or police can do about the reliability of their witnesses, despite any incentives that might be present.

This is a compelling argument, but it ignores several of the problems outlined earlier in respect of expert testimony of any kind. There is no compulsion on courts of law to admit expert testimony,[24]

[24] Although it is worth noting that expert testimony has been admitted in particular instances on the grounds that to exclude it would have the consequence of excluding evidence of known probative value (Loftus & Ketcham, 1991).

and several psychologists have argued that the popular legal distinction between adjudicative and legal fact further hinders the admission of such testimony (Monahan & Walker, 1988).

It suffices for our purpose in this chapter to leave the question as noted, but unresolved: the debate about expert testimony on eyewitness research rages on in the literature (Egeth, 1993), and it is perhaps a bit grandiose to entertain solving it here.

Determinants of application[25]

I wish to re-iterate here that it is not acceptable to discard the notion of an 'applied psychology', even though the way in which it is ordinarily conceptualized is unsatisfactory. In the absence of a solution to this problem, I will make a few observations that serve as a guide to the way in which I approach the particular problem at the centre of the dissertation.

The observations concern what determines whether research gets applied or not. One useful exploration is to systematically examine examples of research that do get applied and to compare them to research that fails to get applied. One such attempt exists in the applied behavioural science literature (Stolz, 1981), which I wish to consider briefly.

Stolz (1981), after a scrutiny of innovations from applied behavioural research, isolated the following determining variables in the successful dissemination and application of research.

1. Research data showed that the innovation was effective.
2. The technology met the continuing mission of the adopting agency.
3. The potential adopter had a pressing management problem.
4. The availability of the dissemination to the potential adopter was timely.
5. Potential adopters were able to view ongoing model and programs.
6. The adoption was proposed by policy makers, rather than by the researchers who developed the technology.
7. The intervention was tailored to local conditions.
8. Those who would have to implement the program were involved in the preliminary research and in asking for the adoption. (Stolz 1981, p 498-99).

[25] I approach 'application' in this section - and perhaps throughout the chapter - from a somewhat narrow perspective, namely as putting knowledge to immediate use. There is a wider perspective one can take here, which is to argue that social science research intersects and creates discourses that operate at a social level, which is surely a form of 'application' of knowledge. This is the sense in which Freud's psychosexual theory has surely changed the entire understanding of childhood development for several generations of people.

There are several other attempts to identify the determinants of application, from a number of diverse disciplinary areas, and these generally support Stolz's conclusions. Stroh (1991) identifies the following five characteristics as being strongly associated with project adoption in the light-construction industry:

> Perceived economic advantage; compatibility; complexity; trialability; observability

Similarly, Luukkonen & Stahle (1990) review the impact of evaluations conducted in four Nordic countries, and suggest that the following characteristics lead to high impact:

1. The evaluation must answer a need for information, and it must be carried out in a methodologically credible fashion.
2. The results must be communicated effectively to those in decision-making positions.
3. The recommendations must help supporters of pertinent positions.
4. Pertinent questions must have powerful supporters.

There seem to be a few lessons here: what is required for the execution of successful applied research is its insertion into an appropriate infrastructure. 'Application' requires a certain jurisdiction: new laws are promulgated by legal government, a change in the policy of a company toward its employees derives from executive decision, and so on. Applications are made by policy makers.

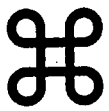
Consequently, applied research must be seen as inherently inter-disciplinary, and the nature of the applied research should reflect the concerns of the discipline or public area that it enters:[26] in the present case, the problem of the identification parade must be understood legally before effective applied research can be conducted. Framing the research solely in terms of the entering research discipline jeopardizes it from the outset.

Conclusion

In this chapter, I have argued that research on identification parades stems from a recent 'drive to application'. However, there are important issues underlying the notion of 'application', and these are frequently unexplored or misunderstood in the eyewitness literature. An adequate treatment of these issues is beyond the scope of the present work, but I have made certain observations that will guide later chapters.

[26] There is an obvious danger here, namely that the research discipline is completely subjected to the concerns and desires of the disciplinary area it is entering, that it becomes a lackey to authority. It may in addition be very difficult to persuade the alien discipline of this danger. These are problems that must be addressed on a case-to-case basis.

In particular, psychological research on eyewitness identifications cannot be justified on the mere premise that it addresses an important practical problem. It must be shown that the research can contribute to the solution of the problem, or the alleviation of the problem. This presumes a sound analysis of the problem, and since it is a legal problem, we must investigate the legal nature of the problem.[27] This is the task for the next chapter.



[27] Of course, the legal problem may be intractable, or there may simply be very little consensus on the nature of the problem. The extent of the investigation will probably be a matter of degree, as it can hardly expect to be a legal tract in its own right.

Chapter 3

Identification Parades in South African Law

Introduction

The identification parade is widely regarded as an important safeguard to an accused who is implicated by identification evidence. I aim to show in this chapter that less protection is accorded to an accused by a 'regularly conducted' identification parade than is commonly supposed. This claim is substantiated by a comprehensive review of South African law on identification parades, and by a brief examination of police practice in South Africa.[28] A particular weakness that I identify concerns the physical resemblance of parade foils: the 'sufficient physical similarity' of foils is set out as a necessary condition for fair parade construction, but, in practice, courts are unable to assess whether this condition has been met. I will use this observation as partial justification for empirical work reported in later chapters.

The review of South African identification law presented here is rather lengthy, but this is in accord with the conclusions reached in the previous chapter. I argued there that applied research on identification problems needs to be thoroughly and carefully informed of the legal character of such problems.

Evidence of identification is by its very nature unreliable, and legal history is replete with examples of convictions which were based on the testimony of honest witnesses who were nevertheless mistaken.[29] Indeed, the English court of criminal appeal owes its existence in large part to the mistaken identifications in the notorious and tragic case of Adolf Beck.[30]

---

[28] Since this chapter attempts a legal analysis, it adheres to a style typical in legal publishing. Footnoting and referencing, in particular, differ from conventional formats in the psychological literature.

[29] As early as 1932 Borchard compiled a volume containing several dozen cases of injustice caused by such evidence (E.M. Borchard, *Convicting the innocent: errors of criminal justice*. New Haven: Yale University Press. 1932). Wills, in *Principles of circumstantial evidence*, 7th ed., at 193 also lists a number of cases. South African courts have certainly been alert to injustices that result from mistaken identifications: in *R. v. Masemang*, 1950 (2) (SA) 488 (A), van den Heever, J.A. at 493, remarks that they "...fill one with apprehension".

[30] Loreburn, Earl. 'Cases of mistaken identity.' *SALJ* 1917 152; Devlin, Lord Patrick. Foreword to 'Identification evidence', J.W. Shepherd, H.D. Ellis, G.M. Davies. Aberdeen University Press, 1982. The creation of this court, of course, did not eliminate the miscarriages of justice that issue from mistaken identifications. In England, there have been three Royal commissions of inquiry this century into miscarriages of justice that issued from identification evidence, two of which postdated the creation of the appeal court. See Devlin, footnote 34 of this chapter, for details.

In the *South African Law Journal*, several articles have acknowledged the inherent dangers of identification evidence. The 'Justice of the peace' writes, in 1926:[31]

> ... mistaken identity is the most likely and common cause of miscarriages of justice, and such miscarriages not only shock the public conscience but give rise to doubt and uneasiness as to the administration of justice. (At 287).

Similarly, the 11th report of the English Criminal Law Revision committee says:

> We regard mistaken identification as by far the greatest cause of actual or possible wrong convictions.[32]

There are several reasons for the unreliability of evidence of identification. The witness engages in a task which is complex and error prone, even in favourable circumstances: he[33] must accurately see and hear a transitory sequence of events, and at some much later stage correctly recall what he witnessed. In addition, the circumstances under which the observation occurs are usually not favourable: the witness probably catches only a fleeting glimpse of the criminal, the illumination of the physical site is poor, and there is a long delay between the event and the occasion on which the witness is required to testify. These conditions militate against accurate testimony, but more threatening even is the fact that the nature of the testimony severely hampers attempts at cross examination. Lord Devlin puts this point lucidly:

> [Identification evidence] ... is exceptionally difficult to assess. It is impervious to the usual tests. The two ways of testing a witness are by the nature of his story - is it probable and coherent? - and by his demeanour - does he appear to be honest and reliable? . . . ..[In] identification evidence there is no story; the issue rests on a single piece of observation. The state of the light, the point of observation, and the distance from the object are useful if they can show that the witness must be using his imagination. But otherwise, where there is a credible and confident assertion, they are of little use in evaluating it. Demeanour in general is quite useless. . . . If a man thinks he is a good memoriser and is not that fact will not show itself in his demeanour. . . (At 76). [34,35]

The courts recognise the danger posed by a bald statement of identification,[36] and approach identification evidence with caution. Numerous judicial *caveats* have been sounded, recommending that identification evidence be thoroughly tested *intra curially* before conviction.

---

[31] 'Identification'. *South African Law Journal* 1926, 287. (No author is given).

[32] Cited in Devlin, 1976, at 76. See footnote 34 of this chapter.

[33] In this, and later chapters, pronouns that typically indicate gender (e.g. he, she) are used interchangeably, and often without the intention of referring to a particular gender.

[34] 'Report to the Secretary of State for the Home Department of the Departmental Committee on the Evidence of Identification in Criminal Cases'. Chairman: Rt.Hon. Lord Devlin. London; Her Majesty's Stationery Office; 1976.

[35] The accused faces an additional danger: the trier of fact might be tempted to conclude that the witness is correct simply because he has withstood cross examination. See *Kola v. R.* 49 (1) P.H.100.

[36] *R. v. Shekelele and another* 1953 (1) (SA) 636 (T), *S. v. Mehlape* 1963 (2) (SA) 29 (A), and many other cases express the need for caution. Indeed, Didcott, J. in *S. v. Ngcobo* 1986 (1) (SA) 905 (N), at 906 says:

The defining feature of cases involving identification evidence is the act of identification, which will invariably occur when the witness is asked whether the perpetrator of the crime is present in court.[37] This will be the identification that the court relies on to convict, but will be preceded by previous acts of identification, which will be made as a matter of course during the police investigation. The most obvious and rudimentary method of pre-trial identification is the staged confrontation: the accused is displayed to the witness, who is asked to positively identify the suspect.[38] This is a very suggestive procedure,[39] and is regarded universally[40] as providing evidence of little value in establishing the true identity of the perpetrator. The patent dangers emanating from one-on-one confrontations provoked judicial criticism, and in response the identification parade evolved, and today is perhaps the most widely favoured test of identification evidence. Williams and Hammelman[41] assert that this is certainly the case in the United Kingdom:

> They [identification parades] have come to be widely considered a valuable and reliable way of providing evidence of identity where this may become an issue before the court. (At 481).

The identification parade has acquired the status of 'most preferred safeguard' against the danger(s) of identification evidence, not only in South Africa, but in most countries. There certainly are other safeguards, notably those set out in the landmark case of *R v. Shekelele*,[42] and I will discuss these, but our focus is on the identification parade.

The history of identification parades

The identification parade appears to have become a regular part of police practice as late as the latter part of the nineteenth century. Lord Devlin traces its origins to a Police Order issued by the Metropolitan police in 1860, following upon "some remarks made by an assistant Judge of the Middlesex session".[43] Isolated parades may, of course, have been held earlier. The test that the identification parade embodies has an intuitive appeal, and earlier societies may well have used a

---

> The danger of mistaken identification is one to which all judicial officers are or should certainly be alive. Enough has been said about it over the years, and in various parts of the world to see to that.

[37] In *R. v. Mputing*, 1960 (1) (SA) 785 (T), Boshoff, J. makes the point that such an identification has little value since the compromising position that the accused is in suggests his identity as the perpetrator to the witness. Hoffman and Zefertt, in their book *South African Law of Evidence*, 1989, Durban: Butterworths, at 615, suggest that the witness would indeed look foolish if he pointed anywhere other than the dock. It is surprising, and should send a chilling warning of the perfidious nature of identification evidence that this, nevertheless, frequently occurs.

[38] Cases do arise where the police are, through force of circumstance, obliged to embark on acts of identification which are less than ideal. See *S. v. Ngcobo* 1986 (1) (SA) 905 (N) for an example.

[39] *R. v. Mputing*, 1960 (1) (SA) 785 (T).

[40] See Shepherd et al., footnote 30 of this chapter.

[41] G. Williams and H.A. Hammelman, 'Identification parades', in Criminal Law Review, 1963, 479-482, 545-555.

[42] 1953 (1) (SA) 636 (T).

[43] Devlin, Ibid. At 112.

similar device. However, it appears to have flourished in legal systems using the English system of prosecution - Lord Devlin points out that it is largely absent in countries that use other systems.[44]

It is not clear when the parade became a regular feature of South African police work. The earliest South African cases I have been able to find that specifically mention the practice, and which lay down 'rules' for its conduct, are *R.* v *Olia* and *R.* v. *Mkize*.[45] The first two editions of the well known South African criminal procedure textbook, by Gardiner and Lansdown[46] make no mention of the practice. In the third edition of 1936 the identification parade is mentioned explicitly for the first time.[47] However, the practice is referred to in two early articles in the *South African Law Journal*.[48]

The presentation of identification evidence in court

Identification evidence can take several forms. It is worth tracing the way that such evidence is presented in court, since this will help to clarify the specific form of identification evidence represented by the result of an identification parade.

Evidence of identification can be direct or circumstantial in nature. I am concerned in this thesis only with direct kinds. Such forms of identification evidence are presented in court by the prosecution. The witness is asked to point out the person who committed the deed that the case has bearing on. This will be the suspect, in the dock. This identification has little value on its own, and serves simply as a reminder that the witness will testify that the accused is the same person she observed commit the deed. Dock identifications have been relied on as primary evidence of identification, but consistent judicial criticism[49] has effectively made this a rare occurrence.

During the course of the trial, the prosecution will lead evidence that the witness has identified the suspect as the perpetrator of the crime in question. This identification may have occurred in a number of ways, but will most likely have been secured at an identification parade, since such identifications are generally held to be most acceptable. Other procedures that have produced eyewitness identifications, which have been admitted in South African courts, include 'showups' or

---

[44] Devlin, op. cit. p 3.

[45] *R.* v. *Olia* 1935 (SA) 213 (T); *Mkize* v. *R.* 1932 (1) (PH) H17 (N).

[46] F.G. Gardiner and C.W.H. Lansdown, *South African Criminal Law and Procedure*. Cape Town: Juta.

[47] Of course, there are earlier English cases that address the practice, and these have been followed from time to time in South African law. See *R.v. Chapman* 1911 7 Cr. App. Rep. 53, *R. v. Palmer* 10 Cr. App. Rep. 77, *R. v. Williams* 1912 8 Cr App. Rep. 84.

[48] Loreburn, *supra*, in 1914; and 'Justice of the peace', *supra*, in 1926.

[49] In the English case of *R. v. Chapman* 1911 7 Cr. f Rep. 53, it was held that the confrontation of witness and accused, in order to secure an identification, is neither fair nor justified. Dock identifications, which are a form of such confrontation, therefore have very little value. Indeed, in *R. v. Palmer* 10 Cr. App. Rep. 77 such a form of identification was held to be sufficient ground for setting aside a conviction.

'confrontations', 'voice identification parades', 'photographic identification parades', and 'dog scent parades'.[50] It is perhaps useful to characterise these procedures as involving an independent test of the witness's ability to identify the suspect. This characteristic distinguishes such identification evidence from a notable and frequent alternative form of identification evidence, where the suspect is well known to the witness, and where such a test of identification ability would at best be superfluous. I concern myself in this chapter with the former class of evidence.

Rules

There are many regulations and guidelines in South African law governing the conduct and evaluation of evidence secured from identification parades. I hope to show that these depend ultimately on the requirement that parade members bear a satisfactory degree of physical resemblance to each other.

Statutes

Section 37(1) of the Criminal Procedure Act No. 51 of 1977 gives police officials the authority to conduct identification procedures. The act states:

> "Any police official may
>
> a) ...
>
> b) make [a person arrested upon any charge, or any such person released on bail or on warning under section 72], available or cause such person to be made available for identification in such condition, position or apparel as the police official may determine..."

The effect of s. 37(1)(b) is that police officials are given the power to conduct identification parades. Arrested persons may be placed on parade (either those still in custody, or those out on bail), but there is no power to compel non-suspects to act as parade 'foils'. Nor do judicial officers have the power to order that a police official conduct an identification parade: in *S. v. Burns*[51] it was held that if a parade has not been conducted the judicial officer's competence is limited to drawing inferences from the failure to have done so.

It is worth noting that the type of identification procedure is not specified by the act (i.e. an identification procedure other than an identification parade could be used), nor are any requirements specified for the satisfactory conduct of the identification procedure. The conduct of the

---

[50] But note that identifications secured at 'dog scent' parades are not considered direct evidence of identification. I include them here because of the similarity that they bear to other forms of identification parade.

[51] *S. v. Burns and another* 1988 (3) (SA) 366 (C).

identification parade (or other identification procedure) is therefore at the discretion of the police official.

A question that naturally arises is whether an accused person is obliged to participate in an identification parade. The Act appears to empower police officials to conduct the parade, if necessary, against the will of the accused. In this respect English law is different, since accused people are able to refuse to participate,[52] but at the expense of forcing the prosecution to resort to another form of identification.

A person placed on an identification parade has the right to counsel. This is assumed in *S.* v. *Jija and others*,[53] and is also the practice in England and the United States of America.

General rules

In order to understand the rules and guidelines that courts have set down in respect of identification parades, it is necessary to consider more broadly the way in which identification evidence is approached. This is because identification parades are usually treated as a test of identification, and the testing of identification evidence is prescribed under a number of so-called 'cautionary rules'.[54]

The cautionary rule with respect to single witnesses

The earliest (and in many cases, logically prior) cautionary rule is that set out in the case of *R. v. Mokoena*.[55] Evidence given by a single witness should be treated with extreme caution. It was held that although it is possible to convict solely on the basis of the testimony given by a single witness, that witness must be honest and creditable, and the testimony must be satisfactory in every material respect. Many cases have followed and elaborated this judgement. Thus, in *Woji v. Santam*[56] it was held that this cautionary rule is also applicable in civil cases, bearing in mind the difference in the onus, namely 'on the balance of probabilities', rather than 'beyond reasonable doubt'. Further

---

[52] See *Phipson on Evidence*, 12th. ed. 1976, London: Sweet & Maxwell, at §1254. An interesting example of how such a refusal was achieved is given in C. Williams, *Identification parades*, Criminal Law Review, 1955. He reports that in the case, *R. v. Lamb*, heard at the Bristol Assizes in 1946 (which is unreported), the defendant muttered "Go on, I'm your man, why waste time". The police officer conducting the parade dissolved the parade and merely asked the witness whether he recognised the prisoner. This identification was upheld at the trial, but the author of the article contends that it was a dangerous route for a court to follow, even in the circumstances of that case.

[53] *S.* v. *Jija and others* 1991 (2) (SA) 52 (E).

[54] Although it is not clear that they can be called 'rules'. See G.L. Peiris, 'Corroboration of the evidence of complainants' *Acta Juridica*, 1981, 139.

[55] *R. v. Mokoena* 1932 (1) (PH) H51. The case is unfortunately not reported in much detail.

[56] *Woji v. Santam* 1980 (2) (SA) 971 (SECLD).

caution is suggested when the single witness is also the complainant - in *R. v. Segoale*[57] it was said that the court must be satisfied that the story belonging to that person whom the onus rested on to discharge must be true and the other false - and when the defence is attempted by an uneducated defendant.[58] Examples of what the courts have held to be testimony that is not satisfactory in every material respect include i) when the opportunity for observation is not good,[59] and ii) when the evidence is vague, or the witness has a motive to give false testimony.[60]

It is mistaken, however, to think that cautionary rules like that formulated in *Mokoena* change the nature of the onus. In *R. v. J.*[61] it was noted that the cautionary rule should not be taken to replace the usual test of proof beyond reasonable doubt. It is only a guide, which may be useful to decide whether the onus has been discharged. In *S. v. Mehlape*[62] the court held that if, at the end of a criminal trial, there is any reasonable possibility of mistaken identity, the state has not proved its case beyond reasonable doubt. Where identity of the offender is the sole issue, the approach to be adopted in weighing up evidence should depend on the circumstances of the case. For example, in *S. v. Khumalo*,[63] it was held that where the accused's story or alibi is not clearly or necessarily false, even if not accepted, in a case of identification this must be seen as hindering the discharge of the onus.

The cautionary rule with respect to identification evidence

The cautionary rule in *Mokoena* is specifically concerned with the testimony of single witnesses. It is relevant to the question of identification evidence insofar as a single witness produces such evidence. A second cautionary rule with particular application to identification evidence was formulated by Dowling, J. in the much cited case *R v Shekele*.[64] It is worth repeating a substantial portion of the judgement in that case.

> Questions of identification are always difficult. That is why such extreme care is always exercised in the holding of identification parades - to prevent the slightest hint reaching the witness of the identity of the suspect. An acquaintance with the history of criminal trials reveals that gross injustices are not infrequently done through honest but mistaken identifications. People often resemble each other. Strangers are sometimes mistaken for old acquaintances. In all cases that turn on identification the greatest care should be taken to test the evidence. Witnesses should be asked by what features, marks, or indications they identify

---

[57] *R. v. Segoale* (SA) 641 (T).

[58] *Filipi v. R.* 1960 (2) (PH) H206 (FSC).

[59] *R. v. Mokoena* 1932 (1) (PH) H51.

[60] *R. v. Ditshego* 1932 OPD 164.

[61] *R. v. J.* 1966 (1) (SA) 88 (A).

[62] *S. v. Mehlape* 1963 (2) (SA) 29 (A).

[63] *Khumalo v. S.* 1964 (1) (PH) H80 (N).

[64] *R. v. Shekelele and another* 1953 (1) (SA) 636 (T). Although judgement was delivered in this case in 1947, it was not reported until 1953. In the period between 1947 and 1953 it was frequently referred to, in its unpublished form.

> the person whom they claim to recognise. Questions relating to his height, build, complexion, what clothing he was wearing and so on should be put. A bald statement that the accused is the person who committed the crime is not enough. Such a statement, unexplored, untested and uninvestigated, leaves the door wide open for the possibility of mistake. Where the accused is an ignorant native who is unrepresented by counsel or attorney and who is therefore unable himself to probe the evidence of identification and where the prosecutor has not done so, the Court should undertake this task, as otherwise grave injustice may be done. (At 638).

The central element of the cautionary approach recommended in *Shekelele* is that identification evidence by an eyewitness should not be accepted unless it has been tested. This approach has been adopted with approval in numerous cases.[65] It is true that courts had acknowledged the need for caution in several earlier cases - in *Kote v. R.*,[66] Kotze, J. suggested that this was obvious to every judicial officer, and in *R. v. Pietersen*[67] a conviction was set aside because evidence of identification had not been tested by the court *a quo* - but the specific requirement set out in *Shekelele* that identifications be tested by directing questions to the witness has been very influential.

The kinds of questions that are put to the witness will depend on the circumstances of the case. Where the witness and the accused are well acquainted, for example, matters of clothing, facial characteristics, and identifying bodily marks are of less importance. The degree of previous acquaintanceship, and the opportunity for observation, will need to be tested.[68] It is probably impossible to specify beforehand all the factors that must be investigated by putting questions to an identifying witness,[69] since each case will have different circumstances.

The greatest care should be taken to test identification evidence, and a witness may be tested in cross examination by requiring him to describe again the appearance of the person(s) he purports to identify. Where the advocate representing an accused was effectively restrained from doing so on the grounds that the court was satisfied as to the credibility of the witness, on appeal it was held that identification evidence must be both credible and reliable,[70] and that the line of cross examination should have been permitted. The court must be satisfied that an identifying witness is honest,[71] but the fact that a witness is *bona fide* and honest is not enough.[72,73] The requirement in *Shekelele* that

---

[65] See, among others, *R. v. Matsha* 1958 (2) (PH) H254 (E), *Pelwan v. S.* 1963 (2) (PH) H237 (T), *R. v. Alluverino* 1963 (4) (SA) 727 (SR), *Poopedi en ander v. S.* 1966 (2) (PH) H407 (T), *S. v. Grosch* 1984 (1) (PH) H53 (SWA).

[66] *Kote v. R.*. 1906 (SA) 189 (E).

[67] *R. v. Pietersen* 1941 (2) (PH) H252 (C).

[68] *R. v. Dladla and others* 1962 (1) (SA) 307 (A).

[69] *S. v. Mehlape* 1963 (2) (SA) 29 (A).

[70] *S. v. Pretorius* 1991 (2) (SACR) 601 (A). A similar position was adopted, earlier, in *R. v. Nomtwana and others* 1961 (4) (SA) 174 (E).

[71] *R. v. Sebeso and others* 1943 (SA) 196 (A), *R. v. Hlatywayo* 1953 (1) (PH) H74 (T).

[72] *S. v. Mehlape* 1963 (2) (SA) 29 (A). The point is made very clearly by Clifford Sully: "Far fewer have been condemned on perjured evidence than on false evidence given in good faith." (Cited in 'Identification', *SALJ*, 1926, at 289).

evidence of identification be tested has been taken by some courts to have strict force. In *R. v. Medupe*,[74] a failure of the court *a quo* to apply the test led the immediate court of appeal to overturn the conviction. Similar decisions were made in *R. v. Motsagie*,[75] *R. v. Alluverino*,[76] and S. *v. Mazibuku*.[77] In addition, evidence of identification should be rigorously tested - in *S. v. Fritz*,[78] the trial judge expressed satisfaction that the identification evidence had been thoroughly tested, but the Appellate Division disagreed, saying that the identification had not been thoroughly enough explored, and it accordingly set aside the conviction.

Several courts have expressed reservations about the test set out in *Shekelele*. In *R. v. Mputing*,[79] Boshoff, J. pointed out that a witness may well recognise a perpetrator accurately, despite not being able to offer any distinguishing circumstances or characteristics. Identification is often a sub-conscious process. This was also said, more emphatically, by van den Heever, J.A. in *R. v. Kumalo*[80] - a description might bear little relation to the perpetrator, and yet the witness may correctly identify the perpetrator. More recently it has been held that the inability of a witness to describe a person identified by that witness is not necessarily fatal to the question of whether the identification is proper.[81] These statements do not purport to disagree with the prescription that identification evidence should be tested, they merely point to the vagaries of such evidence.

In *Mputing*, it was also pointed out that the value of reported characteristics or identifying features is related to how common they are in the population: the less common, the more value such evidence will have.[82] In *Mavuso and another v. The King*,[83] for example, the court held that the accused's voice had a very distinctive ring and tone, and made for reliable identification. In *R. v. T.*,[84] the description given by the witness was said to be true of very many people, and that this reduced the value of the identification.

---

[73] Certainty is not a guarantee of accuracy. Indeed, it may on occasion suggest the opposite, as in the cases of *S. v. Ntsane* 1966 (2) (PH) H408 (N), and *R. v. Weimers and others* 1960 (3) (SA) 508 (A), where the witnesses became more and more certain during their testimony, for no discernible reason.

[74] *R. v. Medupe* 1957 (1) (PH) H64 (GWLD).

[75] 1959 (1) (PH) H42 (GWPD).

[76] 1963 (4) (SA) 727 (SR).

[77] *S. v. Mazibuku* 1966 (2) (PH) H326 (O).

[78] *S. v. Fritz* 1980 (1) (PH) H17 (A).

[79] *R. v. Mputing*, 1960 (1) (SA) 785 (T).

[80] *R. v. Kumalo* 1948 (2) (PH) H200 (A).

[81] *S v. Pretorius* 1991 (2) (SACR) 601 (A).

[82] But, in *Filipi v. R.* 1960 (2) (PH) H206 (FSC), the Court noted that there is a danger attached to distinctive physical appearance - that very distinctiveness alone may induce the court to rely on the witness's identification.

[83] *Mavuso and another v. The King* 1969 (2) (PH) H168 (Swaziland High Court).

[84] *R. v. T.* 1958 (2) (SA) 676 (A).

Perhaps the most important qualification to the cautionary rule set out in *Shekelele* is made in the cases *R. v. Hlatywayo*[85] and *S. v. Poopedi en ander.*[86] Evidence of identification will often be tested in court, and it must be borne in mind that the accused may well be right in front of the accused at the time that this evidence is tested. While it is true that the features whereby a witness recognises someone should be tested, this will have little value if the witness is in the dock and visible to the witness. The witness's description will have more weight if it is committed before seeing the defendant.[87] In *R. v. Y.*[88] Bresler, J. went further, and said that in cases where questions of identification are in issue, details of identification should be investigated at the very earliest opportunity. Otherwise, an accused person might find himself suddenly confronted at the trial with an unexpected mass of important details in regard to the identification. If sufficient details are not furnished to the police, the police should seek greater particularity from the deponents.

However, to test identification evidence with this very important qualification in mind will often imply that evidence of earlier consistent descriptions must be led. A further witness may need to be called to testify to an eyewitness's original description, as happened in the cases of *R. v. Mack*, *S. v. Seanego* and *R. v. Rassool.*[89] This may run foul of the strictures against hearsay evidence, and evidence of previous consistent testimony.[90] The English case *R. v. Christie*[91] for example, ruled such evidence inadmissible. In *R. v. M.*[92] it was held that consistency of description as a test of identification must conform to the rule that a witness cannot be corroborated on the basis of other consistent testimony. Wigmore[93] presents a comprehensive discussion of the issue, and reserves strong wording for this kind of approach in identification cases:

> That some modern courts are on record for rejecting such evidence is a telling illustration of the power of a technical rule of thumb to paralyze the judicial nerves of natural reasoning (at §1130).

---

[85] *R. v. Hlatywayo* 1953 (1) (PH) H74 (T).

[86] *Poopedi en ander v. S.* 1966 (2) (PH) H407 (T).

[87] In the Devlin report previously referred to, Lord Devlin made the recommendation that only the initial description of the perpetrator given by the witness to the police should be admitted as evidence.

[88] *R. v. Y. and another* 1959 (2) (SA) 116 (W).

[89] *R. v. Mack* 1969 (4) (SA) 53 (R), *S. v. Seanego* 1978 (2) (PH) H121 (A), *R. v. Rassool* 1932 (SA) 112 (NPD).

[90] The argument could be made that all forms of pre-trial identification are subject to these strictures, including identifications secured at identification parades. Indeed, in *Rassool*, the court considered this, and held that identifications out of court are admissible as evidence of the identity of the accused as the perpetrator.

[91] *R. v. Christie* 1914 A.C. 545 (H.L.).

[92] *R. v. M.* 1959 (1) (SA) 434 (A).

[93] Wigmore, J.H. *Wigmore on Evidence*. 6th ed. Boston: Little Brown and Company, 1940.

The South African cases that consider this point appear to have decided in favour of allowing such evidence in identification cases.[94] In *R. v. Mack* it was held that although a witness is not normally entitled to reveal the details of a report made by himself which was not made in the presence of the accused, there may be circumstances in which the report would be relevant and therefore admissible. Such evidence could be led, for example, if the circumstances were such that the identification at the parade was challenged. In *S. v. Seanego*[95] it was decided that the report of an identifying witness to a third party may be admissible if it addresses the veracity of the identification, and in *R. v. Velekaze*[96] that evidence of a previous identification is admissible, even if the witness himself does not testify directly to this (but is not admissible when the witness denies the previous identification, as happened in that case). The point was put more generally in *R. v. G.*,[97] where it was said that the Court was entitled to hear evidence of the manner and circumstances in which identifications took place, as the story of this may be important information. In that case, the witness, who was also the complainant, took down the first three letters of a motor vehicle registration number, in which vehicle she had allegedly been raped, and she reported the number to her employer, who confirmed this. While on patrol with a police officer some time afterwards, she pointed the car out, where it stood in a camping ground. This car belonged to the appellant. Still later, the witness saw the perpetrator with another man in a second car, and she took down this number, and reported the matter again to her employer. The police established that the second car belonged to the appellant's brother-in-law. Finally, the witness identified the appellant at an identification parade. The court ruled that the details of this 'story' were important in evaluating the evidence of identification.

The cautionary approach taken to identification evidence is not exclusively concerned with the kind of test recommended in *Shekelele*. It has been said on more than one occasion that the Court should consider corroborative evidence in identification trials very carefully, especially when it is led in favour of the accused. Thus, in *Jubela Necobo v. R.*,[98] it was said if identity is in issue, the fact that a motive cannot be found is of the greatest importance. Similarly, an alibi defence merits careful consideration, and there is no onus on the accused to prove the alibi, it is for the state to show that it is false.[99] If the accused's testimony is shown in some respects to be false, and the accused a liar in this

---

[94] But note that Hoffman and Zefertt (*South African Law of Evidence*, Durban: Butterworths, 1989) are less certain about the status of this evidence in South African courts. Schmidt, in *Bewysreg*, 2nd ed., Durban: Butterworths, 1980, at 368, suggests that evidence of previous consistent testimony is admissible, but only insofar as it details the earlier identification. Where such evidence of identification goes further and details the deeds committed by the alleged perpetrator, the further details are not admissible, except where this is unavoidable.

[95] 1978 (2) (PH) H121 (A).

[96] *R. v. Velekaze* 1947 (1) (SA) 162 (WLD).

[97] *R. v. G.* 1956 (2) (PH) H266 (A).

[98] *Jubela Necobo v. R.* 1926 (1) (PH) H18 (A).

[99] *Mokoena v. R.* 1958 (1) (PH) H99 (A).

respect, it will not be enough to secure the conviction. There may well be good reasons for the false testimony: in *R. v. Motsagie*,[100] it appeared that the accused lied about his whereabouts on a particular night in order to hide the fact that he had broken a curfew, and the untruth was thus not taken into consideration against him.

English law has long battled with the question of corroboration in cases that turn on identity. There have been three Royal commissions of enquiry into identification evidence this century, and in the last of these the principal recommendation was to make convictions in identification cases dependent on corroborative evidence.[101] A similar recommendation was made by a barrister in *R. v. Williams*.[102] In both instances, the recommendation was not adopted, although a less exacting version of the principle was adopted in *R. v. Turnbull*.[103]

The cautionary procedure when witness and accused are acquainted

In *Shekelele*, the witness and the perpetrator were strangers, and the test set out in that case in respect of characteristic or identifying features should not be applied to situations where the accused and the witness are well acquainted. It is not useful to probe evidence of identification for special identifying features when the accused is known to the witness, since the witness may resort to recollection of prior knowledge in answering such questions.[104] It has been held that where the identifying witness knows the identified person well and is honest, her assertion of identity, in good circumstances for observation, is entitled to great weight.[105] In such circumstances even a bald identification may be acceptable.[106] This does not mean, however, that the need to test evidence of identification falls

---

[100] *R. v. Motsagie* 1959 (1) (PH) H42 (GWPD).

[101] Devlin, at 150. The exact wording is important, though:

> We recommend that the trial judge should be required by statute
>
> a. to direct the jury that it is not safe to convict upon eye-witness evidence unless the circumstances of the identification are exceptional or the eye-witness evidence is supported by substantial evidence of another sort; and
>
> b. to indicate to the jury the circumstances, if any, which they might regard as exceptional and the evidence, if any, which they might regard as supporting the identification; and
>
> c. if he is unable to indicate either such circumstances or such evidence, to direct the jury to return a verdict of not guilty. (§8.4, at 150)

[102] *R. v. Williams* [1956] Crim. L. R. 833.

[103] *R. v. Turnbull* [1976] Crim. L. R. 567 (C.A.). In this judgement, it was recommended that judges should draw the attention of juries to the quality and nature of identification evidence in a trial. Where the identification is of 'poor quality', for example, where the witness caught only a 'fleeting glimpse' of the perpetrator, there will be insufficient basis for a prosectution.

[104] *Teka v. R.* 1960 (1) (PH) 171 (C).

[105] *Mtetwa v. R.* 1954 (2) (PH) H199 (A).

[106] *R. v. Hlatywayo*, 1953 (1) (PH) H74 (T). James, J in *R. v. Dladla and others* 1962 (1) (SA) 307 (A), pointed out that acquaintance with the accused must increase the probability that an accurate identification has been made. In *S. v. Gantsho en ander* 1978 (1) (PH) H68 (A) a corollary to this was said to be probably true: namely that it must reduce the likelihood of a mistaken identification.

away,[107] but that the test should take a different form. In particular, the degree of acquaintanceship must be tested - how often has the witness seen the accused?, when did the witness last see the accused?, has the witness ever seen the accused close-up?, has the witness ever spoken to the accused?[108] Matters of clothing, facial characteristics, and identifying bodily marks are of less importance in such cases. What is important is the degree of previous knowledge, and the opportunity for observation, with regard to the circumstances in which it was made.[109] In *R. v. Dladla* it was held that even if the witness makes a mistake concerning the name of the accused, if the witness knows the perpetrator by sight, this may be enough. Some courts have taken the requirement that claims of acquaintanceship should be tested to be a strong one: In *S. v. Mehlape*,[110] the failure to test acquaintanceship was one of two grounds on which a conviction in a lower court was overturned.

Of course, it is not only in cases where witness and accused are acquainted that the opportunities for observation should be very carefully tested. This should also be done where witness and accused are not acquainted. The manner in which the opportunities for observation are tested will depend, as always, on the particular circumstances of the cases. Thus, in *S. v. Nhlabali*,[111] the witness saw the perpetrator, briefly, by the light of an acetylene blowtorch and the reflected light of two battery torches aimed at a safe, and this was judged inadequate opportunity. In *R. v. T.*[112], the opportunity for observation that a rape victim had in a dark park, at a fair distance from the nearest light, was held to be adequate by the trial court, but the appeal court ruled that the opportunity was affected by the fact that victim and perpetrator were complete strangers, and this probably hindered the opportunities to too great an extent to allow accurate observation.

The cautionary rules in respect of identification evidence should of course not be adopted to the detriment of common sense: the probative value of an identification depends on the circumstances of the case, and each case must be judged on its merits.[113] In *Van Rensburg v S.*,[114] it was ruled that although the trial court had not followed the prescribed tests, the witness had observed the perpetrator when he took his mask off, and since she said she recognised the perpetrator as someone whom she knew, she was unlikely to be mistaken, given that the opportunities for observation were good.

---

[107] The fact of acquaintance does not in itself show that the witness is not mistaken. *Sibiya and others v. R.* 1956 (1) (PH) H136 (A).

[108] *S. v. Mehlape* 1963 (2) (SA) 29 (A).

[109] *R. v. Dladla and others* 1962 (1) (SA) 307 (A).

[110] 1963 (2) (SA) 29 (A).

[111] *S. v. Nhlabali* 1967 (2) (PH) H304 (A).

[112] 1958 (2) (SA) 676 (A).

[113] *R. v. Masemang* 1950 (2) (SA) 488 (A).

[114] *Van Rensburg v. S.* 1968 (2) (PH) H329 (A).

Somewhat more questionably, in the case of *S. v. Nango*,[115] where the opportunities for observation were constrained, and a policeman testified to the identity of a man who attempted murder on a second policeman with an axe, the Court held that the witness had special powers of observation (*spesiale observasievermoëns)* - by virtue of his training and membership of a special police unit - and that this made the identification more likely to be accurate.[116]

Many cases have stressed the need to consider evidence of identification in sight of the totality of the facts and the circumstances of the case, and I have provided examples of this type of approach above. What is perhaps not very clear is how this affects the inherent weighting to be given identification evidence. In *Mputing*, Boshoff, J. contended that an identification must be reliable enough to stand on its own (*selfstandig*), but in other cases it appears that fairly weak identification evidence has been accepted when the circumstances are such as to suggest the guilt of the defendant. This is especially true of cases involving identification parade evidence, and I will consider the implications in this regard at some length later in the chapter.

Identification parades are an important form of test under the cautionary approach recommended in the several landmark identification cases. This is explicitly acknowledged in *Shekele* and *S. v. Gordon*,[117] and indeed, as we will see in the next section, they are widely considered indispensable to the pre-trial identification process.

Identification parade rules

I have not been able to trace the 'originary moment' of the identification parade in South African police or legal practice (if, indeed, such a moment exists), and the absence of discussion of the practice in the South African law reports before 1932 is noticeable. It appearance probably owes much to police practice in England, and to the early cases of the English court of criminal appeal.

One of the earliest of these cases, *R. v. Chapman*,[118] spoke strongly against the practice of 'confrontations'. It was said that confrontations of witness and accused - in order to secure identifications - are neither fair nor justified. Dock identifications, in particular, have very little

[115] *S. v. Nango* 1990 (2) (SACR) 450 (A).

[116] An earlier case, though, held that the fact that a police officer of proved integrity and reliability as a witness makes an identification does not render it more reliable. It is irrelevant to the question of whether the officer is mistaken in his identification. *Madiba v. R.* 1947 (2) (PH) H272 (N).

[117] *Gordon v. S.* 1970 (1) (PH) H68 (A).

[118] *R. v. Chapman* 1911 7 Cr. App. Rep. 53.

value.[119] The usual and proper way to conduct identifications is to place the suspected man with a sufficient number of others, and to have the identifying person pick out a man without assistance.[120] It is wrong to point out the suspected person and ask "Is that the man"? At the time of this case, decisions made by the English appeal court were binding on South African courts, and *Chapman*'s case is frequently cited in the law reports, albeit not immediate to its reporting.

The prescription for pre-trial identification in *Chapman* is echoed in South African cases that bear on the matter. Thus, in *Mputing*, it was said that where there is uncertainty regarding the identity of the person who committed the misdeed, an identification parade should be held. Similarly, in *Kola v. R.*,[121] Schreiner, J.A. held that it was of the greatest importance that identification parades should be held, except where they are useless (e.g. where the witness knows the defendant well). He further stated that it is unsatisfactory to rely on an identification of a witness not acquainted with the accused without it being tested by a parade.

A stronger case could hardly be made for the practice of identification parades than that made in *Kola* and *Mputing*. Courts have certainly relied on the purported reliability that a parade identification imparts - in *S. v. B.*,[122] for instance, where a woman was raped by two men, her identification of these at a parade was the only evidence in front of the court, and they were convicted on this basis. Nevertheless, many judges and legal scholars have been fully mindful of the dangers that attend parades - in *Kola* itself, Schreiner issued the following lucid warning:

> But an identification parade, though it ought to be a most important aid to the administration of justice, may become a grave source of danger if it creates an impression which is false as to the capacity of the witness to identify the accused without the aid of his compromising position in the dock. Unsatisfactory as it may be to rely upon the evidence of identification given by a witness not well acquainted with the accused, if that witness has not been tested by means of a parade, it is worse to rely upon a witness whose evidence carries with it the hall-mark of such a test if in fact the hall-mark is spurious. (At 169).

How does a court decide whether this 'hall-mark' has been cast truly, or is counterfeit? In the cases, this question has been approached in terms of the notion of 'regularity': an identification at a regularly conducted parade can be accorded great weight, but irregularities render such identifications of little weight.[123] This very important notion, however, has not been carefully defined in judicial

---

[119] van den Heever, J.A., in *R. v. Kumalo* 1948 (2) (PH) H200 (A), was even more forthright. He called such evidence "worthless", and said that it was "calculated to endanger the liberty of innocent persons" (At 331). In *R. v. Madubedube* 1958 (1) (SA) 276 (O) it was pointed out that it does not only endanger innocent suspects, but also lets the guilty go free.

[120] Similarly, in *R. v. Williams* 1912 8 Cr App. Rep. 84, it was held that for the purposes of identification, the suspected person should not be presented alone.

[121] *Kola v. R.*. 1949 (1) (PH) H100 (A). Many other cases have acknowledged the important role of identification parades as a form of pre-trial identification. See *Mtetwa v. R* 1954 (2) (PH) H199 (A); *R. v. Madubedube* 1958 (1) (SA) 276 (O).

[122] *S. v. B. and another* 1980 (2) (SA) 946 (A).

[123] *R. v. Masemang*, 1950 (2) (SA) 488 (A).

discussion in South Africa, and is approached most frequently by identifying specific 'irregularities'. If a parade suffers from no discernible irregularities, it is considered to be regularly conducted. Some courts have offered more general heuristic advice: in *R. v. Olia,* for example, it was said that an identification parade should be conducted in such a manner that there can be no doubt whatsoever as to the genuineness of the identification.[124] Nevertheless, most of the case law concentrates on specific sources of irregularity, and I will consider these in some detail.

In *Kola* it was held that the onus is on the state to exclude all possibilities that an irregularity might have occurred in a parade. This position has been adopted and applied, with approval, in many cases.[125] In *Kola* itself, the Crown was not able to show that the identifying witness could not have seen into the parade room while the identification parade was being formed, and could therefore not eliminate the possibility that the identity of the suspect was made known (however inadvertently) to the witness. In *S. v. de Bruin,*[126] it was not clear whether the accused was the only person on the parade with a dark suit, and since the State did not lead evidence to show that this was not the case, it did not exclude the possibility that an irregularity had occurred.

The word 'irregularity' is perhaps not very well chosen, since it invites comparison with the way the term is used in Acts like no. 31 of 1917 (section 370). In R. *v. Sebeso*[127] as much was argued by counsel for the defendant: specifically, it was said that certain irregularities which occurred at an identification parade prejudiced the trial *per se*, and should have been considered irregularities in terms of the aforementioned Act. This contention was dismissed by the Court, and it was held that irregularities at identification parades are not irregularities at the trial, but merely factors bearing on the value of the identification evidence adduced there. This ruling was affirmed in *R. v. W.*[128] The notion of an 'irregularity' in the conduct of an identification parade is perhaps best exemplified by the well known remarks of van den Heever, J.A. in *Masemang*, at 493:

> But where such identification rests upon the testimony of a single witness and the accused was identified at a parade which was admittedly conducted in a manner which did not guarantee the standard of fairness observed in the recognised procedure, but was calculated to prejudice the accused, such evidence standing alone can have little weight.

---

[124] It is not clear that this is very helpful. Later in the thesis I will argue that it may be more accurate to think of identification parades as always leaving some doubt as to the accuracy of the identification.

[125] See *S. v. Burgess* 1978 (1) (PH) H10 (N), and *S. v. Mlati* 1984 (4) (SA) 629 (A).

[126] *S. v. de Bruin* 1967 (2) (PH) H325 (A).

[127] *R. v. Sebeso and others* 1943 (SA) 196 (A).

[128] *R. v. W.* 1947 (2) (SA) 708 (A). In *R. v. Y. and another* 1959 (2) (SA) 116 (W), it was said that even as a failure to observe the regulations in an identification parade is not an irregularity at the trial, such a failure will not be overlooked.

An irregularity at a parade is an aspect of the conduct or structure of the parade that prejudices the accused (and a regularly conducted parade is, conversely, one that does not in any way prejudice the accused). In *Mlati*, Botha, J. made this clear:

> Die uitdrukking "calculated to prejudice the accused" sinspeel natuurlik nie op 'n doelbewuste beplanning van die polisiebeamptes om die beskuldigde te benadeel nie; dit verwys na die bestaan van feite wat, objektief beoordeel, tot gevolg het dat die beskuldigde blootgestel is aan benadeling, hoe ook al daardie feite tot stand gekom het. (At 635)
>
> (*Translation.* The expression "calculated to prejudice the accused" naturally does not refer to a premeditated intention on the part of the police officials to compromise the accused; it refers to the existence of facts which, objectively considered, have the effect of compromising the accused, however those facts came into being).

I will return a little later in the chapter to the question of the effect that parade irregularities have on the way identification evidence is evaluated.

Irregularities in the constitution of the parade

The principle that appeared to lead English courts to reject identifications secured at 'showups', or confrontations, was that these procedures are extremely suggestive. In *Chapman's* case,[129] it was contended that the correct procedure is to place the accused among a sufficient number of other people, and have the witness attempt an identification. The exact number which is sufficient was not considered by that court, but in South Africa several legal scholars have said that it is eight,[130] although no justification is given for this particular number. The parade members chosen to serve as parade 'foils' should be similar in physical appearance to the accused: in *R. v. Sebeso*,[131] similarity of colour, build and clothing was considered adequate, and in *R. v. Weldon*,[132] it was said that utmost care must be taken in arranging identification parades to see to it that clothes of the parade members will not suggest to a witness that the suspect must be sought amongst those parade members who are differently dressed.[133] Irregularities in terms of this last requirement have frequently led to the need for higher courts to overturn decisions. In *Masemang*, the accused was dressed in a jersey which had a distinctive dark maroon colour, whereas the other parade members had on jerseys a shade lighter. This was considered an irregularity. Similarly, in *Mlati*, the defendant was the only parade member who wore a black leather jacket at the identification parade, and he was immediately identified by the

---

[129] 1911 7 Cr. App. Rep. 53.

[130] Hoffman and Zefertt, at 616; V.G. Hiemstra, *Strafprosesreg*, Durban: Butterworths, 1967, at 73, say that a total of eight people (including the suspect), is required. This has been a requirement under English law for some time (see G. Williams and H.A. Hammelman, *Identification parades*, in Criminal Law Review, 1963, at 479), and may well be the source of these claims.

[131] *R. v. Sebeso and others* 1943 (SA) 196 (A).

[132] *R. v. Weldon* 1947 (1) (PH) H39 (A).

complainant, who had earlier reported that one of her assailants was wearing such a jacket at the time of the offence. In this case it was held that although the relevant section in the Criminal Procedure Act of 1977 gave a police official the power to place a suspect on an identification parade in any apparel the official might determine, the official should attempt to ensure that an identification at a parade would have probative value, and this might well entail ensuring that the manner in which the accused is dressed is not unduly suggestive. The State argued that in the instant case the accused had been asked before the parade if he was satisfied with the parade, and he had replied in the affirmative, but Botha, J. concluded that there was no onus on the accused to ensure that the identification parade was regularly conducted, but indeed that the onus rested on the State. In *S. v. Sibanda*,[134] it was said very clearly that a police official must use his discretion appropriately when considering the apparel to place identification parade members in - distinctive clothing may 'assist' the identification to the detriment of the accused.

A parade should be constituted by people of similar physical appearance, but this requirement cannot be absolute, and will usually be met in degree. Thus, in *R. v. Tusi & another*,[135] the defendant had a beard of fairly short length, and it appeared that the beards of other parade members varied somewhat from this. This disparity was not, however, considered sufficient to warrant any concern over the reliability of the identification.

In some cases it will be very difficult indeed to construct a parade which consists of people reasonably similar to each other: in *Weldon*, Scheiner, J.A. asserted that this is probably the case when the defendant is a woman. Perhaps the most extreme example of this problem is to be found in the case of *S. v. Seanego*,[136] where the defendant had a green mark on his head. In this case, the trial judge found it understandable that the police had not attempted to construct an identification parade. In practice, the police find it very difficult to find people of suitable physical appearance who are willing to serve as parade foils,[137] and this task is still more difficult where the accused has a peculiar identifying feature.[138] In several discussions I had with police officers in the SAPS (South African Police Service) charged with arranging identification parades, it transpired that a common strategy for

---

[133] See also *Molife v. R.*. 1955 (1) (PH) H62 (T).

[134] *S. v. Sibanda and others* 1969 (2) (SA) 345 (T).

[135] *R. v. Tusi and another* 1957 (4) (SA) 553 (N).

[136] *S. v. Seanego* 1978 (2) (PH) H121 (A).

[137] "It is all very well for the new police orders to say that the accused "will be placed among a number of persons, as far as possible of similar age, height, general appearance and class of life as himself or herself", but the practical difficulties in the way of doing so must be immense, while anything like a real failure to carry out the order almost entirely destroys the efficacy of the parade". (Justice of the peace, 'Identification', *SALJ*, 1926, at 288).

[138] One approach to this problem is to place a covering on the feature, and to put a similar covering in the same place on each of the foils. This happened in *Molife v. R.* 1955 (1) (PH) H62 (T), where the accused had a bandage around his head. The practice is apparently commonplace in the U.S.A. See also the case of the one eyed girl who allegedly robbed a sailor, *SALJ*, 1926 at 288.

obtaining foils is to recruit policemen not involved in the case at hand, and government employees from neighbouring buildings. It is worth noting that this practice is strongly discouraged in England, and has largely fallen away there. The reasons appear to be[139] i) police officers serving as foils may know who the suspect is, and convey this unwittingly to the witness; ii) in country districts most witnesses would know all the resident police officers, and they would therefore be useless foils; iii) police officers often have a distinct bearing and manner, and this will mark them as different from the suspect.

Nevertheless, that the parade should consist of people of reasonable physical similarity is an important requirement. In this respect, the evaluative criteria set by the courts are only partially helpful. Colour, build and clothing may be important attributes on which to determine similarity, but they are not sufficient, nor are they adequate. Facial similarity is perhaps a far more important attribute for parade members to have.[140] Little has been said about this attribute, although it could be argued that the requirement of physical similarity *ipso facto* includes facial similarity. Nevertheless, a trial court is in a disadvantaged position when it comes to evaluating facial similarity, since it will frequently have no useful record of the constitution of the parade. Nowadays police often make a photographic record of the parade, and may submit such record to the Court, but they have not always done this, and there is no compulsion on them to do so. Even if a photographic record of the parade is available for inspection, it is not clear that it takes the matter much further, for how is one to evaluate facial similarity? This, of course, is the theme for the empirical work reported in later chapters.

Parades with multiple accused members

The practice of identification parades, as commonly understood, involves the insertion of an accused into a line of other, innocent people. This is frequently not the case in practice; it is perhaps as common to encounter parades with multiple accused members as it is to encounter single suspect parades. There is some evidence to suggest that the former type of parade is inherently inferior to the latter, which I will detail in a later chapter. It is notable that it is not a requirement in South African law that a parade should contain only one suspect. However, where a number of persons are under suspicion and an identification parade is held, the parade should not consist only of the suspected persons, even if this means that more than one parade must be held.[141] This is not a strong requirement - in the very case in which it was formulated (*Wildman en andere v. S.*), it was held that a

[139] See Williams and Hammelman, footnote 41 of this chapter.

[140] I will substantiate this claim in greater detail when I discuss psychological research on identification parades, namely in the following chapter.

[141] *Wildman en andere v. S.* 1968 (2) (PH) H356 (A).

parade which consists only of suspects could yield an identification that is not totally valueless, although it will be greatly reduced.

Irregularities in the procedure of the parade

The Courts have frequently made recommendations regarding the procedural aspects of identification parades. The concern is with how the parade is conducted, and the purpose of the recommendations is to avoid situations in which there is suggestion as to the identity of the accused, or even the possibility that there is such suggestion.

Thus, in *R. v. Mkize*,[142] it was held that identification parades should be arranged so that suspects and witnesses do not come into contact with each other before the parade. In that case a conviction was set aside because the accused people and the witnesses had arrived independently at a police camp and had been in a position to come into contact without police supervision. Witnesses should also not be in a position where they can see into the parade room while the parade is being formed, and they should also not see any of the parade members before the parade is conducted.[143]

A similar restriction is applicable to groups of witnesses. When witnesses are assembled preparatory to an identification parade, police officials ought to prevent them as far as possible from engaging in conversation which concerns the identity of the person sought. A member of the service, where possible, should be present to see that such an instruction is not infringed.[144]

There should be no opportunity at all that allows the suggestion of the identity of the suspect to the witness(es), no matter that this might only happen inadvertently. Courts have gone so far as to say that the witness must not know that the police have a definite suspect in mind, and that therefore, the police official conducting the parade should not know the identity of the suspect in the parade.[145]

In *R. v. Nara Sammy*, several of these rules were affirmed. The presiding judge rejected an identification made at a parade on the grounds that i) witnesses were herded into a common room before the identification and discussed the appearance of the perpetrator; ii) the officer conducting the parade had seen the parade being assembled and therefore knew who the suspect was and could have communicated it to witnesses, even if this was an unlikely event; and iii) the presiding officer failed

---

[142] *Mkize v. R.* 1932 (1) (PH) H17 (N). This was the earliest reference I could find to the term 'identification parade' in a report of a South African case.

[143] *Kola v. R.* 1949 (1) (PH) H100 (A).

[144] *R. v. Weldon* 1947 (1) (PH) H39 (A).

[145] *R. v. Nara Sammy* 1956 (4) (SA) 629 (T), also *R. v. Y. and another* 1959 (2) (SA) 116 (W). In *R. v. Masemang*, 1950 (2) (SA) 488 (A), a police officer accompanied the witness on the way to the parade, and from their reported conversation he appeared to intimate that he knew the identity of the suspect. This was considered inappropriate and irregular.

to warn of the possible absence of the offender, and therefore probably implied that he was present. For the latter reason, the judicial officer recommended that the instruction "if such person is present on the parade" should be incorporated into police instructions to witnesses at identification parades. This has been taken in some cases to be a strict requirement. Thus, in the Transkei case reported at appeal as *S. v. Mpopo*,[146] the witness was asked by a police sergeant "to point out the person whom she suspected to have killed the deceased". The sergeant's error received a sharp comment from the trial Judge, Munnik, C.J.:

> I think you must go right back to Sterkspruit and let them teach you how to run an identification parade again. You have just messed up another murder case. How could you go and ask a witness who does she suspect of murdering somebody? (At 428)

However, in other cases the absence of such a warning has not been treated with the same gravity. In *S. v. Hay*,[147] where such an instruction was not issued, it was held that the omission of the instruction '*indien hy op die parade aanwesig is*' (tr. If he is present in the parade) need not necessarily compromise the accused.

The disclosure of a parade identification in court

The Courts have also made several observations and recommendations regarding the reporting of parade identifications at the trial. In the first place, it has been stated that only the witness who made the identification should be permitted to testify to this fact in court. In *Molife v. R.*,[148] the crown led evidence that the accused had been identified at two of three identification parades, but failed to call the identifying witness in respect of the first identification. The court dismissed this evidence.

If an identification parade was held, and a person who could well have identified the perpetrator did not identify the accused at a parade, it is the duty of the prosecutor to bring this evidence to the attention of the court, especially when identification is the only or main issue.[149] Indeed, the failure of witnesses who could be expected to identify the perpetrator must be evidence in favour of the accused - it must imply some doubt as to the evidence that the accused is the perpetrator.[150] In *Nksatlala v. R.*, Schreiner, J.A. held that the test to be applied when giving weight to this failure is whether the failure to identify raises a reasonable doubt in the mind of the court.

---

[146] *S. v. Mpopo* 1978 (2) (SA) 424 (A).

[147] *Hay v. S.* 1970 (1) (PH) H98 (A). A similar position was taken in *R. v. Y. and another* 1959 (2) (SA) 116 (W).

[148] *Molife v. R.* 1955 (1) (PH) H62 (T).

[149] *Molife v. R.*, supra.

[150] *S. v. Molakang* 1979 (2) (PH) H154 (O).

The way in which the identification is made at the parade appears also to be of some importance. Thus, in *R. v. Hlatywayo*,[151] Steyn, J. said that an unhesitating identification at an identification parade increases the believability of the identification, while a hesitating identification suggests the converse. This assertion, however, is not entirely consistent with the many observations made regarding witness demeanour in identification cases. Thus, in *Mehlape*, Williamson, J.A. noted

> The often patent honesty, sincerity and conviction of an identifying witness remains, however, a snare to the judicial officer who does not constantly remind himself of the necessity of dissipating any danger of error in such evidence. (At 32).

Earlier, in *Masemang*, van den Heever, J.A. issued a similar warning:

> The positive assurance with which a witness will sometimes swear to the identity of an accused person is in itself no guarantee of the correctness of that identification. (At 493).

In police practice, there appears to be a requirement that the witness physically touch the person she wishes to identify at the parade. I can find no discussion of this in South African cases, but police officials are of the opinion that such a ruling has come from the courts. It is a particularly unpopular requirement among rape complainants, for obvious reasons. In England, it has long been a requirement that the witness touch the identified person,[152] except where the witness is very nervous.

The details of how an identification was made at an identification parade may thus be taken into account at the trial, but an issue that has arisen in this regard is the status of the police record of such a parade. (Later in the chapter I will discuss the mechanism by which police keep records of identification parade, namely by requiring police officials to complete form SAP 329). Is such a record privileged, and if so, which type of privilege is it protected under?

The Courts are divided on this issue. In *S. v. Mpetha*,[153] Williamson, J. ruled that identification parade documents are privileged, but in *S. v. Jija and others*[154] the Appellate Division formed a different opinion. There it was held that a police record of an identification parade is not a privileged document, either in the sense of legal professional privilege, or in the sense of witness statement privilege: it is difficult to see how privilege could attach to a document completed in the presence of the accused and their legal representatives, who are entitled to question the correctness of the information set out in the document. Such a document is a contemporaneous note of the *res gestae* of the parade. Alternatively, the court held, if there is such privilege, judicial discretion should be

---

[151] *R. v. Hlatywayo* 1953 (1) (PH) H74 (T).

[152] The Home office has issued recommendations and guidelines on the conduct of identification parades from time to time, and in the earliest of these that I have been able to trace (the 1925-1926 guidelines, reproduced in Shepherd et al. [see footnote 30], at 123), the touching requirement is included at the reported insistence of the Secretary of State.

[153] *S. v. Mpetha and others* 1982 (2) (SA) 253 (C).

exercised in cases where the document becomes relevant and it should then be made available to the defence.

There is also divided opinion in the cases concerning the effect of previous acquaintance on the probative value of identifications made at parades. In *Kola*, at 169, Schreiner, J.A. said that "identification parades ... would be useless ... where the identifying witness knows the suspect apart from the occasion of the crime." This proposition appears straightforward, and a rationale for it may be found in the cases which consider the question of acquaintanceship more generally. There is little point in testing the identification of a witness who is acquainted with the accused, since the witness may rely on previous knowledge of the accused[155] - that is to say, on knowledge that does not stem from her observations at the time of the event. The test will be easily passed, and this will go nowhere toward the real goal of such a test, which is to determine whether the witness is correct in her assertion that the accused was the perpetrator she saw at the event.

However, in the case of *S. v. Rademeyer*,[156] a different decision was arrived at by Muller, J.A. In this case, two witnesses identified the accused at an identification parade as the man they had seen commit a murder. It appeared that they knew the accused by sight, although both claimed that their knowledge of him was slight. The advocate for the accused argued that their previous knowledge of the accused negated the value of the identifications made at the parade. Muller, J.A. had a different view:

> Ek kan glad nie sien hoe hierdie voorafkennis die bewysgewig van hulle uitkenning op die parade negeer of enigsins affekteer nie, en dit veral waar die verhoorhof bevind het dat hulle eerlike en betroubare getuies was. (At 303).
>
> (*Translation*: I simply cannot see how this prior knowledge negates, or even affects the evidentiary weight of their identifications at the parade, especially since the court *a quo* found that they were honest and reliable witnesses).

Other forms of parade

What I have called the 'identification parade' in this chapter is really only one form of the parade, namely that in the visual modality, and might more appropriately be called the 'visual identification parade'. Just as a claim of identity might rest on the visual observation of particular identifying characteristics, so it may also rest on aural observation, or observation of voice characteristics. It is also a mistake to think that an identification parade need occur corporeally, that is to say as a task in

---

[154] 1991 (2) (SA) 52 (E).

[155] *Teka v. R.* 1960 (1) (PH) 171 (C).

[156] *S. v. Rademeyer en ander* 1980 (2) (PH) H191 (A).

which a witness inspects an array of live people: often, an array of photographs will be presented to the witness, and the pre-trial identification will occur solely from the photographs. In this section of the chapter I examine these alternative forms of identification parade.

Voice parades

Identification on the basis of voice is not a common form of identification evidence led in our courts, but it occurs from time to time, and the courts have considered the merits of this type of evidence. An early English appeal case, *R. v. Keating*,[157] held that it is possible that a sufficient identification can be made on the basis of a person's voice alone, and accordingly admitted such evidence. In the earliest reported South African cases to consider this type of identification evidence, there was some doubt about the admissibility of identifications secured at voice identification parades. In *R. v. Gericke*,[158] the advocate for the accused argued that the relevant section in the Criminal Procedure Act of 1917 did not make provision for making someone participate in a voice parade, since a person's voice could not be said to be a distinguishing feature or mark. Alternatively, the advocate argued, making the accused participate in a voice parade runs foul of the common law, which prohibits forcing people into self-incriminating actions (i.e. the tenet *nemo tenetur se ipsum prodere* has application).[159] These arguments were overruled: the court declared that there was no difference in principle between a visual identification parade and a voice identification parade. In a later case, *S. v. M* 1963,[160] it was specifically ruled that evidence of voice identification is admissible as a voice cannot fail to be viewed as a characteristic or distinguishing mark. In addition, the argument regarding self-incrimination was rejected, since this particular stricture has application to forced confessions, and the intention behind the stricture is to guard against evidence made unreliable by the way it is adduced. Identification parades are not confessions, and are therefore not protected under the tenet.

Several cases have made specific recommendations regarding the conduct of voice identification parades. Thus, in *Gericke*, it was said that i) there should be more than four voices in the parade; ii) some of the parade members' voices should be known to the witness - not merely one but more than one; iii) the witness should not be taken to the suspect last, since there is no-one else to identify at

---

[157] *R. v. Keating* 1909 2 Cr. App. Rep. 61.

[158] *R. v. Gericke* 1941 (3) 211 (C). In this particular case, the voice parade was conducted by putting five people including the accused in different cells, and having the ear-witness listen to a police officer speak to each in turn, without seeing any of them.

[159] Of course, this argument can also be used against visual identification parades. This has occurred in both English and American courts (see Williams and Hammelman, footnote 41 of this chapter), but it is not accepted there. See Nathan R. Sobel 'Eyewitness identification: Legal and practical problems', New York: Clark Boardman, 1991, at 8-1, for a comprehensive discussion of this issue in U.S. law.

[160] *S. v. M.* 1963 (3) (SA) 183 (T).

that point. No justification is offered in support of recommendation i), nor is any offered for the peculiar recommendation in ii).[161]

In *R. v. Chitate*,[162] Quénet, J.P. applied the reasoning from *Shekelele* and other identification cases to the whole issue of voice identification. The same care must be taken with aural parades as is taken with visual identification parades, since the danger of mistaken identification is also present. Questions must be put as to what features of the voice (e.g. timbre, loudness) are recognised. The opportunities for correct aural observation should also be tested (e.g. the loudness of the perpetrator's voice, the number of people speaking at that time). The voice test must occur as early as possible,[163] and certainly before the witness has had an opportunity to hear the accused speak. 'Voice foils' who participate in the parade must be chosen to resemble the suspect's voice. The whole purpose of such tests is to discover whether a witness's conclusion is reliable, and such reliability will remain untested if the test itself was improperly conducted. In *S. v. M.* 1972[164] some of these recommendations were confirmed, and several added. In particular, it was said that witnesses must not know who the accused is, and they should be separated from each other to counteract suggestion from one to another.

However, in some cases the notion of a voice parade has been severely questioned. In the Swaziland High court, in *Mavuso and others v. The King*,[165] the presiding judge considered voice identification parades to be of dubious value, since the suspect could disguise his voice, or he could choose to say little, or nothing. In *S. v. M.* 1963, where a voice identification parade was not held, but appeared applicable, the judge questioned the possibility of satisfactorily conducting a voice identification parade, noting that he had grave doubts "whether any voice identification parade would not culminate in pure frustration." (At 183).

Photograph parades

In several countries, it is now common for police to use photograph identification parades, or 'photospreads' as they are known in the U.S.A. This practice is not yet common in South Africa, although evidence of identification from photographs has been admitted.

---

[161] Indeed, the latter recommendation would enable the witness to disregard one or more of the voices in the parade, and the size of the parade would then possibly be smaller than that recommended!

[162] *R. v. Chitate* 1966 (2) (SA) 690 (RA).

[163] Not as in that case, where the test occurred after the accused had testified in court for some time.

[164] *S. v. M.* 1972 (4) (SA) 361 (T).

[165] *Mavuso and another v. The King* 1969 (2) (PH) H168 (Swazi).

Again, an early precedent can be found in the English appeal court cases. In *R. v. Allen Ferguson*,[166] it was held that photographs of suspected persons, whose arrest is under consideration, may be shown by the police for purposes of recognition. However, persons recognising suspects should not afterwards be asked to identify them at the trial. If they are, all the facts of the recognition should be disclosed. In *R. v. Melany*[167] a similar ruling was made, but it was added that identification by photograph should be conducted where corporeal identification is not possible, and that this identification should not precede a corporeal identification.[168]

These authorities were cited in the Southern Rhodesian case of *R. v. Jackson*,[169] where a woman was arrested on the basis of an identification from a single photograph, and tried for cheque fraud.[170] Here it was ruled that a series of photographs should be used if police are to use photographs for purposes of identification. If a corporeal identification parade is held after the photograph parade, the value of an identification made there is greatly lessened, but this effect is less severe if a series of photographs is used for the parade, with a photograph of the accused among them.

In a recent case, *S. v. Shandu*,[171] Mr. Justice Didcott spoke against the value of photograph parades. He held that the conditions under which they are conducted can scarcely duplicate the conditions of a corporeal identification parade. Resemblance of parade members is not guaranteed in a photospread, and there are no settled procedures and regulations governing the technique to protect the accused from influence and suggestion. Where such a parade is held, the reliability of an identification rests to a large extent on the number of photographs exhibited, and the fairness with which they are presented. It is not clear however, that the difficulties the learned judge identifies are either unique to photograph parades, or insurmountable. All identification parades appear to suffer under the requirement of physical resemblance, and indeed, the problem may be greater for corporeal identification parades. In photograph parades, it is possible to collect beforehand a large number of photographs from which to select suitable foils, and parades could be 'standardized' for fairness using empirical techniques. In addition, the safeguards against suggestion currently in effect with respect to

---

[166] *R. v. Allen Ferguson* 1924 18 Cr App. Rep. 145.

[167] *R. v. Melany* 1924 18 Cr App. Rep. 2.

[168] However, the recent Home office guidelines of 1978 in respect of the conduct of identification parades have amended these recommendations somewhat. Here it is suggested that i) any photograph of a suspected person that is shown to a witness be embedded in an array of at least 12 photographs (including that of the suspected person); ii) the photographs in the array should be of a similar type, and the people who have been photographed should be as similar to each other as possible; iii) the witness should be explicitly informed that the photograph of the perpetrator might not be in the array; and iv) a witness who has made a firm identification by photograph should always be asked to attend an identification parade unless it is unnecessary or impracticable for some other reason. (Home Office Circular No. 109 of 1978, reproduced in Shepherd et al. [see footnote 30] at 128).

[169] *R. v. Jackson* 1955 (4) (SA) 85 (SR).

[170] The accused was also placed on an identification parade, though, where she was identified by a witness who had made an identification from a photograph shown her by the police.

corporal identification parades could very easily be enacted for photographic parades. A real difficulty with photograph parades appears to be the difference in quality of identifying information that the witness is given. For instance, a corporeal parade provides more information of a three dimensional nature than provided in photographs. Nevertheless, I will later consider empirical evidence which suggests that the difference in this regard is not great.

It is worth digressing at this point to make the observation that witnesses frequently come into contact with photographs of accused people independently of identification tasks, and these photographs may 'contaminate' the memory of the perpetrator. This criticism applies with equal force to instances where witnesses are shown photographs of suspects during the course of police investigations, and to instances where the press publishes photographs of the suspect before the trial commences (and on occasion, before identification parades have been held). In such instances, a witness may come to believe that the police have arrested the perpetrator, and mistake her original memory of the perpetrator for that she has gleaned from the photograph. Glanville Williams warned the legal community some time ago of the terrible potential for injustice that the uncontrolled police use of photographs creates.[172] In England, the Home Office has heeded his advice (albeit somewhat belatedly), and has issued detailed instructions for the use of these in cases of criminal identification.[168]

Dog scent parades

The final type of identification parade I wish to consider here[173] is one that is both rather unusual, and that has attracted considerable attention in the *South African Law Journal*.[174] I mean the type of identification parade in which the 'witness' is a dog, and the parade is commonly known as a 'dog scent parade'.

I will discuss this form of parade briefly, since the evidence contained in an identification from such a parade is not direct evidence in the way that evidence delivered by an eye- or ear- witness is direct evidence. The dog that makes the identification did not observe the event in question, but attempts to correlate a scent presumed to belong to the perpetrator with a member of an array of people

---

[171] *S. v. Shandu* 1990 (1) (SACR) 80 (N).

[172] See footnote 30.

[173] I do not consider 'item' parades (as conducted in *R. v. W.* 1947 (2) (SA) 708 (A), where a suit and a bicycle were placed among similar suits and bicycles for identification), since Courts have not ruled specifically on these parades, except to admit them as evidence.

[174] See L. H. Hoffmann "Those Dogs Again" 1974 *SA Law Journal* vol 91 237.

assembled for 'scenting'. The recent case of *S. v. Shabalala*,[175] taken on appeal to the Appellate Division carefully considers this form of evidence.

In that case, a man awoke one night to find an intruder attacking his wife. He attempted to defend his wife from the intruder, who fled from the scene of the crime, leaving sand shoes behind him in the flight. These shoes were later given to a specially trained bloodhound, one 'Tilly', who sniffed the sandshoes, and then sniffed each member of an assembled array of people. Tilly stopped at the appellant, barked, and touched him with her paw. This was led as evidence of identification in the court *a quo*, where it was admitted as having evidential value, although not much weight was attached to it. The Appellate Division, however, ruled that such evidence is inadmissible, confirming a judgement made much earlier, in *R. v. Trupedo*.[176] The court did say, though, that it does not follow that *Trupedo* is the final pronouncement on the matter in all circumstances. The principal reason for the exclusion of such evidence is the untrustworthiness of the evidence, and where this element is sufficiently reduced the actions of the dog would become relevant and therefore admissible. Mere proof that the dog comes from special stock and that it has been specially trained in tracking will not suffice. Additional evidence explaining "the faculty by which (these) dogs... are... able to follow the scent of one human being, rejecting the scent of all others" (At 734), is required.[177]

Police practice in South Africa

It has been noted on more than one occasion by the courts that the conduct of identification parades is a matter largely in the hands of the police, and that this therefore gives them a wide discretion.[178] For this reason, it is important to examine the in-house procedures the South African Police Service follows in respect of identification parades.

Police officials appear to receive no (or very little) formal training in the conduct of identification parades, either during their basic training in police college, or during later specialized training. An officer who is frequently charged with organising parades at Caledon Square, Cape Town,[179] told me that the very first experience he had with an identification parade was as an organizing official. Two policemen in the identikit unit at Caledon Square reported similar experiences, and asserted in

---

[175] *S. v. Shabalala* 1986 (4) SA 734 (A).

[176] *R. v. Trupedo* 1920 AD 58.

[177] A recent article summarizes the scientific evidence at hand, and comes to the conclusion that there is inadequate scientific support for trusting canine behaviour in scent identification parades. A conviction based on such evidence, the author argues, is based more on superstition than reason. A.E. Taslitz. 'Does the cold nose know? The unscientific myth of the dog scent lineup'. Hastings Law Journal 1990, vol. 15, 42.

[178] *S. v. Sibanda and others* 1969 (2) (SA) 345 (T).

[179] This was communicated in a telephonic interview (see footnote 21).

addition that police officials are often left to their own devices when constructing parades.[180] There is no 'pool' of foils, for example, that can be drawn on for such a purpose, and police frequently request civil servants working in neighbouring buildings to serve as foils. The following 1989 newspaper report corroborates these claims:

Magistrate says police tried to mislead him[181]

by Mziwakhe Hlangani

A Port Elizabeth magistrate expressed outrage and disgust yesterday at irregularities allegedly committed by policemen during an identification parade. The magistrate, Mr. A W Meiring, said he had found out that policemen had tried to mislead him by covering peep holes through which suspects could be viewed during an identification parade with fingerprint ink. Mr. Meiring said he had never thought a police officer would go out of his way to falsely and deliberately implicate a suspect so that he was sentenced to jail for a long term. The magistrate said he had always trusted policemen and had sentenced thousands to jail because he had relied on police evidence. ... The court had established that there had been irregularities in an identification parade and that a junior officer who was inexperienced had been placed in charge of the parade, said Mr. Meiring. There had been holes in walls though which suspects could be seen at a New Brighton identification parade. Conversation could be clearly heard between walls separating complainants, police and accused. ... Mr. Meiring ordered the inspection in loco on Wednesday morning. He also ordered that nobody should leave the courtroom before he was ready to leave for the inspection. "But somebody apparently sent a message through for the openings to be covered, because when I got there, I found wet fingerprint ink spread over the openings", Mr. Meiring said. ... The magistrate also said it was clear - from the evidence of Constable K A Stemele who was in charge of the parade - that he did not know the identification parade procedures. He said that the incorrectness with which the parade was conducted was illustrated by the fact that none of the policemen had been able to explain why they had four people on the parade when the complainant had to identify three suspects.

This lack of training may be due to a belief on the part of police management that police officers do not require training, that the purpose and nature of identification parades is self-evident. It may, on the other hand, be due to a belief that parades can be adequately conducted and regulated by a printed set of procedures. Indeed, a standard police form governs the administration of identification parades, and police officials are obliged to complete the form whenever they conduct a parade.

The form in question (SAP 329) is reproduced as Appendix A. Since the form appears to be the chief mechanism guiding in-house police practice, it is worth examining in some detail.

There are 34 sections. Sections 1-4 require details concerning the police member in charge of the parade, the police station it is conducted at, the charge(s) against the suspect, and the details of the investigating officer who has issued the instruction that the parade be conducted. Sections 5-9 record details of the suspect(s) to be placed on parade, confirm that suspects were informed of their entitlement to legal representation, and record who the representatives were, if appointed. Sections 10-11 record the names of the photographer and interpreter, if present. Sections 13-16 record the names of police members who guarded and escorted witnesses, before, during, or after the parade. Section 17 requires the police member in charge to indicate how many people were on the parade, and to certify that they were "...of about the same height, build, age and appearance and were dressed more or less similar to the suspect(s)." Sections 18-19 record that suspects were informed of the

[180] This information was elicited in 1987, when I consulted police about an entirely different research matter.

[181] Eastern Province Herald, April 28, 1989, pp 1-2.

charges against them, of the purpose of the parade, and that they were entitled to make any reasonable request(s) in respect of the parade. These sections also record what their requests were, and what steps were taken as a result of the request(s). Sections 20-21 record that suspects were asked if they were satisfied with the parade, their answers to this question, and the names of legal representatives, if present. Section 22 records the name, age and address of all parade members, and section 23 records whether a photograph was taken of the parade. Sections 24 - 25 record i) the procedure of the parade, including the positions taken up by parade members, from left to right, ii) the name and language of the first witness, ii) that the witness was asked to point out the suspect(s), if on the parade, by touching his/her/their shoulder(s), iii) the time taken by the witness to point out a person(s) on the parade, the result of this identification, and the reaction of the witness during the pointing out of person(s). Sections 26 and 27 record that suspects are given the opportunity to change their positions before the next witness is introduced, whether they avail themselves of this, and what the new positions of parade members are. Sections 28-31 record details regarding identifications made by additional witnesses, following the same format as Sections 24-27. The final three sections, 32-34 record additional remarks, and a certification of just report from the conducting officer.

The form is lengthy, and detailed. Many of the guidelines that have come from the courts, as summarised at some length in this chapter, are embodied in the form. It is notable, though, that these guidelines are rarely developed beyond the original wording. In the case of the requirement of physical similarity, no details regarding the physical characteristics of parade members are recorded, and if a photograph is not taken at the time of the parade, the court has no way of assessing whether the requirement has been met, short of issuing summons to the people who served as foils. Some of the guidelines are also not implemented in the form: in particular, the requirement that the officer conducting the parade should not know the identity of the suspect is not enforced.

Police practice appears to depend very heavily, and perhaps unwisely, on SAP 329. I should stress, however, that this conclusion derives from a handful of interviews, and a more comprehensive investigation may show police practice in a more positive light.

Evaluation of South African law on identification parades

In the preceding sections of this chapter I attempted a comprehensive overview of South African law on identification parades, and to a lesser extent, of police practice in respect of such parades. It is time now to provide an evaluation of this law, and to identify points of concern. The problems are in most cases not particular to South African law. In the following chapter, where I discuss psychological research on identification parades, it will be clear that several of the problems are inherent to the task.

The status of identification parade 'rules'

The power of police officials to conduct identification parade is enacted at the level of statute. The courts have frequently noted that the practical administration of parades is accordingly at the discretion of the police service. Nevertheless, courts have made many remarks about the conduct of parades, and, as discussed at some length in earlier parts of this chapter, have also formulated rules by which to evaluate their 'regularity'. But of what force are these rules? Are they 'rules' at all? This is an important question, and I will consider two troubling aspects of the formulation and implementation of parade rules in the case law. In the first place, the rules are applied inconsistently across cases, and secondly, they are applied as a matter of degree, so that it is possible on occasion that most of the rules are broken, but the evidence is still admitted. A few examples from preceding sections will be provided as substantiation.

Courts are inconsistent about the effect that prior acquaintanceship has on the probative value of an identification at a parade. The Appellate Division held on one occasion that prior acquaintanceship renders the value of such an identification questionable, and on another that it does not affect the value of the identification at all (see page 53). Similarly, courts have held that identifications made by policemen are not more likely to be accurate than those made by lay people (see footnote 116), but that they are when made by policemen who belong to certain units, since they have 'special powers of observation' (see page 44). Of course, inconsistencies are to be expected in any body of discourse that evolves through the practice of precedent; more troubling is the way in which parade rules are used along an implicit gradient of application. The requirements of conduct set out in some rules can be compromised, and even flagrantly broken, and there may yet be little effect on the evidentiary value of the identification. Thus, it is possible that a parade may consist of only suspects (see page 49), even though an axiomatic rule is "...to place the suspected man with a sufficient number of others, and to have the identifying person pick out a man without assistance" (see page 45). In one case, it was decided that the qualification 'if the person you saw commit the crime is present' should be added to the standard police lineup instructions (see page 51), but in another case the absence of such an instruction was not accorded any significance (see page 51).

One might be forgiven for concluding that the entire approach of the courts to the rules set by precedent is of this 'sliding nature'. This approach has its genesis in the legal doctrine of case-wise interpretation: each case must be considered on its merits (page 43), and the facts of the case must be considered in their totality and in sight of all the circumstances (page 44). While it is clear that blind

adherence to technical rules may "...paralyze the judicial nerves of natural reasoning",[182] this doctrine makes the 'rules' set down for identification parades indeterminate. They are not rules at all, and the protection they afford falsely accused people is questionable.

The super-ordinate rule which cannot be evaluated

Much of the case law has concentrated on protecting accused people from procedural irregularities. There are also important protections that have to do with how identification parades are constituted, and one of the most important of these is the stipulation that parade members should bear a sufficient degree of physical resemblance to each other. The conceptual basis of this rule is not set out clearly in the cases, and I will attempt in the next section of the chapter to explicate the legal basis for the claim. For the moment, though, the question I wish to address is how courts or police officials evaluate compliance with this rule.

It is clear that police officials labour under the task of finding foils similar to the suspect (see page 48). Courts appear to evaluate this rule in a somewhat *ad hoc* fashion. Thus, similarity of the foils is judged in particular cases by matches of colour, build (height and weight), clothing, and facial hair, but nothing is said about matches in respect of facial similarity (see page 49), nor is an acceptable degree of similarity defined. This is perhaps a near-intractable task, and the courts can therefore hardly be expected to perform it well. Nevertheless, the consequence is that physical similarity is a type of potentially 'confounding' or 'third' variable, whose effects on the fairness and efficacy of the parade are concealed. It throws into question the entire basis and purpose of the identification parade as an evidentiary tool.

What are identification parades for?

Despite the fact that the identification parade has been in use now for over 100 years, its explicit purpose has rarely been carefully examined by courts. Some commentators assert that a lineup primarily provides protection against the suggestiveness inherent in other methods of identification such as a direct face-to-face confrontation between the witness and suspect (Devlin, 1976).[183] But what will the witness's positive identification show in such a situation? Will it show that the witness is reliable, and that the details of his/her testimony can therefore be taken seriously? Or will it

[182] Wigmore, see page 40 of this chapter.

[183] It is worth noting here the recent research by Gonzalez, Ellsworth, and Pembroke (1993), which points out that no-one has ever provided empirical substantiation of this claim. Indeed, in several experiments conducted by Gonzalez et al. showups had a lower degree of response bias than lineups.

provide direct evidence against the suspect, in the sense that it increases the probability that the suspect is guilty? In short, should identifications secured at parades be treated as independent evidence of identity, or should they be treated as tests of assertions of identity?

Both of these aims are built into the structure of the lineup. The witness is asked, if able, to indicate the perpetrator from a number of 'foils', who are known to be innocent of the deed in question. If the witness identifies one of the foils, his/her evidence regarding the identity of the perpetrator is considered less reliable than it would otherwise. The identification parade clearly serves on the one hand as a check on witness reliability. Yet, a positive identification is also taken as evidence against the suspect: indeed, the courts are at great pains to ensure that the identification at a parade constitutes an independent piece of evidence.[184] This would not be the case if the identification parade served merely as a reliability check. Consistency of identification would serve as a measure of reliability in its own right. Instead, the courts prefer to consider identifications as independent evidence regarding the suspect's guilt.

The implications of the two aims are quite different. If the intention is to test the reliability of a claim of identity, then the identification of a suspect at the parade is of marginal value, since it merely serves to corroborate the claim. If the intention is to secure independent evidence of identification, then the parade identification is of great significance. In the next chapter, where I survey the psychological literature on identification parades, I will show how this issue has been addressed from a somewhat different perspective.

This chapter has shown that there are many complications in the treatment given to identification parades in South African law. What I would like to take forward to later chapters as a central argument, is that both courts and police officials are presently unable to assess whether a most important requirement has been met, namely that an identification parade consist of people who are sufficiently alike.

[184] See the remarks made by Boshoff, J. in *Mputing* (discussed on page 44 of this chapter).

Chapter 4

Psychological Approaches to Identification Parades

Introduction

In the previous chapter, I reviewed the legal safeguards against the vagaries of eyewitness evidence, particularly with respect to the practice of identification parades. There I made the point that courts are well aware of the dangers that attend identifications made by eyewitnesses, and I outlined the rules that have evolved in South African courts in response to these dangers. One of the conclusions I drew from that discussion was that one of these 'rules', namely the required physical similarity of suspect and foils, is very difficult to evaluate, and may operate as the legal equivalent of a confounding variable. In the present chapter, I review a body of psychological research on identification parades, and conclude - in part - that variations in physical similarity may indeed affect witness performance in identification parades, but that these effects are not well understood. I also review measures of parade fairness reported in the psycho-legal literature, and consider some of the research aimed at optimizing parade procedure and structure.

There is a long history of psychological interest in legal problems surrounding evidence of identification. As early as the beginning of the 20th century European psychologists were demonstrating the notable imperfections of human witnesses to their students, and were publishing regularly in continental and American journals (see for examples Whipple, 1912; Stern, 1910). However, most of the research on eyewitness issues has accumulated in the last twenty years, and since little of the early research is directly relevant to the question of identification parades, I will focus instead on the more recent psychological literature.

The literature will be discussed under three headings. In the first, I consider research on measures of parade fairness. In the second, I outline psychological conceptualizations of the nature of the identification task, and discuss the relation of these to practical recommendations on foil selection. In the third section, I consider research on structural features of parades.

Research on measures of fairness

The notion of a 'biased identification parade' is so well known in most countries that it is almost part of contemporary folklore. Apart from providing the less concerned with a good belly laugh, the biased parade is a genuine outrage to jurists and observers of the law alike, since it leads to grave injustice. A notorious example of such a parade is cited in Ellison & Buckhout, (1981). Minneapolis

police arrested a black man as the prime suspect in a case they were investigating, and placed him in a parade otherwise constituted only by white men. Similar cases have occurred in South Africa; in the previous chapter I referred to the case of *Pelwan*, where a parade was composed of three Indian men and three white men, the suspects being three Indian men. In *Pelwan* and the case referred to above, the parade was clearly (perhaps self-evidently) biased - and easily identifiable as such - but more usually the bias will be difficult to identify and to describe. In addition, such identification and description will always depend on inherently subjective observations; this is especially the case for infractions of the requirement of physical similarity.

Several psychological studies have addressed the problem of measuring parade fairness, and several of these have suggested interesting measures of fairness. These are worth reviewing in some detail. In Chapter 6 I address these measures again, and attempt to develop ways of reasoning inferentially about them.

The method of the mock witness

The contemporary psychological interest in measures of lineup fairness stems in large part from an article published by Doob & Kirshenbaum (1973), in which they reported the results of an unusual experiment. The experiment tested a police lineup used in a Canadian case, *R v Shatford*, for fairness. The case turned on an identification made by a single eyewitness. The authors questioned that the eyewitness in the case was able to make an identification at a parade, given the fact that the only physical description that she was able to provide to the police was that the perpetrator was 'attractive'. Doob & Kirshenbaum suspected that the witness was basing her identification on, at best, 'partial memory'[185] of the suspect), or at worst, on her own description to the police. In the first stage of the study, 20 subjects who were 'blind' to the identity of the suspect were asked to rate the members of the parade for attractiveness. The suspect received a substantially higher mean attractiveness rating than any of the other foils, suggesting that the suspect was distinctive, and could well be selected by a witness who remembered only that the perpetrator was attractive. In the second stage of the study, Doob & Kirshenbaum showed a photograph of the identification parade to 23 'mock witnesses' (witnesses who had not been present at the original crime), along with the 'description' of the suspect given to police by the witness. They reasoned that subjects who had not been present at the crime should not be able to identify the suspect with a probability exceeding that expected on the basis of random selection (1/number of lineup members = 1/12 = 0.083). 64% of the witnesses correctly

[185] A term coined by Doob & Kirshenbaum.

identified the suspect. Doob & Kirshenbaum reported that the probability of this occurring is less than 0.001, and concluded that the lineup was biased.

Doob & Kirshenbaum suggested that this method - which has come to be known as the method of the mock witness - could be used in general to assess the fairness of identification parades. The method posits that the lineup is fair when the proportion of witnesses choosing the suspect does not exceed that expected on the basis of mere random choice.

Functional and Nominal size

The intention of Doob & Kirshenbaum's study was to provide a measure of lineup fairness. Information regarding lineup fairness is provided to some extent by the proportion of accurate mock witness choices, but it is certainly not complete. An important additional feature of a lineup appears to be the number of plausible foils that it contains. If some members of the lineup are implausible, then the expected proportion of accurate mock witness identifications should perhaps be calculated as 1/k', where k' is the number of plausible foils. A clear example of this reasoning is given in the South African case of *S. v. Pelwan.* In this case, the court dismissed evidence of identification from a lineup whose membership consisted of three Indian suspects and three white foils, where the crime had been committed by three Indian men. The court reasoned that the witness had pointed out the only three people on the parade that he could have pointed out: that is, the parade consisted in effect of only three members.

Wells, Leippe & Ostrom (1979) coined the term 'functional size' to deal with this type of situation, viz. that where the 'nominal' size of the lineup and the number of plausible foils clearly diverge. They suggested a measure of 'functional size', which has the intent of providing an index of the number of plausible lineup members, and is therefore also a measure of lineup fairness. The measure relies on the mock witness task introduced by Doob & Kirshenbaum, but avoids certain problems encountered by Doob & Kirshenbaum's procedure for evaluating fairness when 'null' and 'perfect' foils are present in the lineup.

Doob & Kirshenbaum's measure of fairness is insensitive to null and perfect foils. If a lineup, consisting of 9 foils and the suspect, is fair, the suspect will be identified with probability = 1/10. However, if 5 valueless foils are added to the identification parade (valueless in the sense that they are so unlike the original description that they have an irrelevant probability of selection), this probability will be 1/15, suggesting that the additional foils have decreased the likelihood that the suspect is identified. This is not the case, by definition, and casts some doubt on the value of the

assumption of equiprobability. An argument with a similar conclusion can be made for the case of 'perfect' foils (foils chosen at rates exactly equivalent to chance expectation).

Wells et al. (1979) suggest an alternate measure, which avoids comparing the proportion of accurate identifications[186] to the expected proportion under an assumption of equiprobability. The measure is defined as *D/N* (where *D* is the number of mock witnesses who choose the defendant, and *N* is the total number of mock witnesses), and represents the proportion of mock witnesses who identify the suspect. It is insensitive to the problem of null and perfect foils since it depends only on the frequency with which the defendant is identified, and not on the size of the lineup. Wells et al. (1979) suggest that the measure has a useful interpretation when transformed to its reciprocal (*N/D*), which is the number of functional members of a lineup, hence the term 'functional size'. It is this index that they suggest as a measure of lineup fairness. A lineup is fair when the 'functional size' and the 'nominal size' are identical. In the hypothetical lineups d and e of Table 4.1, which represent lineups with clearly divergent numbers of plausible foils, functional size is 2 and 6 respectively, and this difference corresponds quite well to the apparent difference in fairness of the lineups (i.e. from visual inspection of the array frequencies).

	Lineup Member							
	1	2	3	4*	5	6	Not Present	Functional Size
d	5	3	6	30	9	3	4	2
e	8	9	8	10	9	9	7	6
f	1	3	20	10	21	4	1	6

* = suspect.

Table 4.1 Functional size in a number of hypothetical lineups.

It is doubtful that the statistic suggested by Wells et al. (1979) really provides an index of lineup size. Malpass (1981) argues this quite convincingly, so there is no need to repeat his observations here, except that the measure depends only on the proportion of accurate mock witness identifications, and takes no account of the distribution of identifications across the foils. It is possible for the functional

[186] Several distinctions need to be made in respect of the accuracy of witness identifications at parades, although these depend at some level on whether the witnesses are 'mock witnesses', or 'simulated witnesses', or real witnesses. In parades where the suspect is the perpetrator, a correct identification is the identification of the suspect. All other choices are incorrect. In parades where the suspect is innocent, the correct witness response is to indicate that the perpetrator is not present. Two other responses are possible, but they should be distinguished, at least conceptually. If the witness chooses one of the foils, this is an error, and it will be obvious in all three kinds of parade that this is the case. However, where the witness chooses the innocent suspect, it will not be clear in 'real parades' that the choice is mistaken. It has become conventional in parade research to separate these types of errors. Where the distinction is important, I will adhere to the convention.

size of a lineup to be identical to its nominal size and for the distribution of identifications to exhibit a clearly different picture about the number of plausible foils. Lineup f of Table 4.1 is one example (the functional size of the lineup is 6, suggesting 6 plausible lineup members, but visual inspection suggests that this is an over-estimate),[187] and it is easy to imagine many others.

Effective Size and Defendant Bias

Malpass (1981; Malpass & Devine, 1983), has provided a thorough analysis of the contributions made by Doob & Kirshenbaum (1973) and Wells et al. (1979) to the measurement of lineup fairness. In his analysis he argues for a distinction between *lineup size* and *lineup bias*. Lineup size refers to the number of plausible members that the lineup contains, and it contributes directly to the fairness of the lineup by decreasing the probability that the defendant is identified by a witness who wilfully chooses at random. Lineup bias, on the other hand, is the extent to which mock witnesses choose the defendant at rates greater (or smaller) than chance expectation. Because both these components contribute to the fairness of a lineup, a measure of each is required. Malpass (1981) suggests a measure of each.

The first measure to be considered is what Malpass calls 'effective size', which is intended to measure what Wells et al. thought functional size would measure.

It is clear that an identification parade should have a sufficient number of members to provide some conventionally agreed level of protection against the threat of a witness choosing randomly. Thus, if a parade consists of only two members it provides little protection, since the probability of being identified on the basis of random identification is 0.5. On the other hand, a parade consisting of ten members would reduce this probability to 0.1. It is not clear what identification probability should be regarded as safe: this will depend on how the courts weigh the danger of convicting the innocent (a Type 1 error, in statistical terms) against the danger of setting free the guilty (a Type 2 error). However, most legal systems suggest that parades should have between six and ten members, and so seem prepared to convict between 10 and 17% of all innocent suspects.[188]

Malpass suggests that it would be a mistake to think that an identification parade of some nominal size automatically provides adequate protection against the threats posed by a small parade. This is

[187] This particular distribution of lineup identifications is not as unlikely as it may seem. There are several reported instances of mock witness trials where foils have drawn more identifications than the suspect (cf. Buckhout, Rabinowitz, Alfonso, Kanellis & Anderson, 1988).

[188] Note, though, that this statement has disguised in it the assumption that innocent and guilty suspects are equi-probable, whereas it is presumably the case that guilty suspects are much more likely to be members of identification parades. Also, the truth of the statement depends on the further assumption that witnesses choose randomly, which is unlikely.

because one or more of the lineup members may be implausible, and a witness who has no information at all about the true identity of the perpetrator will be able to disregard the implausible foils and randomly choose the suspect with a probability greater than 1/nominal size. The probability of being randomly selected from a lineup depends not only on the size of the lineup, but also on the quality of the foils (specifically, the number of plausible foils).

In order to evaluate a lineup, Malpass argues, the critical thing to know is how many plausible foils it contains. Although Wells et al.'s (1979) measure of functional size attempts to estimate this quantity, it has several faults. Malpass suggests an alternative to functional size, namely 'effective size', which is the number of effective choice alternatives presented to mock witnesses. In a more precise notation,

$$\text{Effective size} = k_a - \sum_{i=1}^{k_a} \frac{|o_i - e_a|}{2e_a}$$

where o_i = the (observed) number of mock witnesses who choose lineup member i; e_a = the adjusted nominal chance expectation ($N*[1/k_a]$); k_a = the adjusted nominal number of alternatives in the lineup (original number - number of null foils).[189]

The intent of the measure is to reduce the size of the lineup from a (corrected) nominal starting value by the degree to which members are, in sum, chosen below the level of chance expectation. As is clear from the formula, the absolute value of the difference between observed and expected values is taken, and the sum is divided by 2, in order that the calculation reflect the sum of differences where $o_i - e_i < 0$. (Malpass reasons that the important departures are those below chance expectation, since lineup members who fail to draw identifications are poor foils).

The assumption underlying the notion of effective size is an appealing one. One or more of the foils in a lineup may present an inadequate test of a witness who has little more than only very general knowledge of the appearance of the offender, and we shouldn't take the ability of a witness to reject such foils very seriously. The calculation of effective size acts on the assumption by reducing the nominal size of the lineup as a function of the departure of identifications of individual foils from that expected by an equiprobability model. For many distributions of identifications the measure does seem to give an indication of the number of foils that could reasonably be considered present. Lineups g, h, and i in Table 4.2 are clear examples.

[189] The notation used here differs slightly from that given by Malpass (1981), to ensure consistency with later theoretical development, in Chapter 6.

	Lineup Member							
	1	2	3	4*	5	6	E_a	E_b
g	0	25	5	25	3	2	2.83	3.00
h	10	10	9	10	11	10	5.90	5.90
i	1	0	12	12	0	11	3.17	3.17
j	7	7	7	24	8	7	4.60	4.60
k	12	6	9	13	14	6	5.10	5.10
l	6	19	3	20	8	10	4.45	4.45

E_a = effective size calculated with adjustment for null foils. E_b = effective size calculated without adjustment for null foils. * = suspect

Table 4.2 Effective size in a number of hypothetical lineups.

However, the measure also produces estimates that seem intuitively at odds with visual inspection of particular lineup distributions (lineups j, k and l are cases in point). In Chapter 6, when I return to measures of lineup fairness, I will provide further arguments in this regard.

Defendant bias

Malpass (1981) points out that effective size does not provide a measure of bias towards the suspect, but an estimate of the number of plausible foils present in a lineup. Thus, it is possible for a suspect to participate in an unbiased lineup of very small effective size. Imagine that only three members (including the suspect) of a ten member parade are plausible choices. If the suspect is chosen by mock witnesses with a probability of 1/3, there is no bias toward the suspect, despite the large number of redundant foils. We should not reject this lineup because of the differential probability of successfully choosing the suspect, but rather because of the increased risk that the suspect will be chosen at random from the lineup with three effective members (i.e. at a rate of 0.3̇ per identification).

Bias towards (or against) the suspect is another matter. Malpass advises that, in principle, we follow Doob & Kirshenbaum's method, which was described earlier. Here bias is equated with the departure of the proportion of suspect identifications from that expected under an assumption of equiprobability. He suggests that the 'size' of the lineup is better estimated by the calculation of 'effective size', which I discussed at some length immediately above. Thus, defendant bias will be measured by departures of the suspect identification probability from 1/[effective size of lineup]. The idea here is that the likelihood of being selected randomly by a witness is less a function of the mere size of the lineup than a function of the number of plausible foils present in the lineup.

How should the measures of parade fairness be used?

Each of the measures discussed above can be used in a practical sense to evaluate parade fairness. The method in each case is that of the 'mock witness': research subjects are given a description of the

suspect,[190] and they are required to select the parade member who best fits this description. The measures are then calculated on the basis of the mock identifications.

The measures are intriguing, and indeed, useful contributions to the considerable problem of evaluating the quality of identification parade evidence. They have been applied to several real cases, where they have been used to argue for the biased, or fair, nature of identification procedures used there (Doob & Kirshenbaum, 1973; Buckhout, et al., 1988; Brigham & Pfeifer, 1994). In addition, some validation research conducted by Brigham & Brandt (1992) shows a substantial degree of correspondence between judgements of fairness made by law enforcement officials, and estimates of functional and effective size.

The contributions bring a bit of hard-headed empirical thinking to a problem that is usually dismissed as intractable, and not amenable to investigation. However, there are three reservations about the measures that I would like to articulate at this point.

In the first instance, they may be of limited legal utility, particularly in the (South African) legal system used as a reference point in this thesis. The mock witness task is impractical. While it is certainly possible for the defence - or a self-critical police force - to monitor identification parades using some combination of the measures described above, it would be expensive to do this on a regular basis. All of the measures require that a sizeable experiment be conducted, in which identifications are collected from mock witnesses: this entails the recruitment of a substantial number of civilians each time an identification parade is held.[191] The original parade members would also have to be retained for the 'mock' parade. Both of these requirements make for large expenses, and it is unlikely that the police will be able to sustain the required mock witness recruitment for very long. Furthermore, most identification parade cases appearing before the courts in South Africa are defended *pro deo*, so it is unlikely that the defence will be willing to incur the expenses associated with these methods.

In the second place, the methods all proceed indirectly, and therefore fail to directly address the features of the parade that determine its fairness, or bias. If a high proportion of mock witnesses are able to correctly identify the suspect, we do not know what it is about the parade that enables them to do so. It may be that suspect-foil similarity is low, or it might be that the suspect has a particular, distinctive feature, or it may even be a procedural irregularity. Part of the problem is the lack of

---

[190] Typically, the original description provided by the witness. If this is not available, then a description can be generated by the experimenter (eg. the experimenter obtains a description from a subject who is allowed to see the suspect under similar circumstances to those of the original event).

[191] It is not possible to exactly estimate the required number of subjects per parade, since this will vary according to both the size and the fairness of the lineup. However, a six member parade would probably require approximately sixty subjects.

conceptual clarity about what an identification parade is supposed to achieve. I will return to this problem later in the chapter, when I discuss specific strategies for selecting foils.

Finally, a substantial problem with the measures is that they are used without any attention to their statistical properties. This is a significant failing, since the mock witness task has its basis in a probability model: the number of correct mock witness identifications is compared to that expected under an assumption of equiprobabilistic choice. This comparison must explicitly take heed of random variation in choice, or it is bound to capitalize on chance. Similar arguments can be made for the measures of functional and effective size. One of the central tasks that I hope to accomplish in this thesis is the development and application of inferential statistical reasoning to these measures. This work is reported in Chapter 6.

Theories of identification parades

In Chapter 3, I pointed out that although the lineup has been used for at least 100 years, its explicit purpose has rarely been carefully examined by courts. This is also a failing which has characterized identification parade research, at least until recently. There are now at least two or three serious attempts to make clear the conceptual basis of the identification parade, and these have important implications for the selection of parade foils, and for the evaluation of identifications secured at parades. I deal with some of this theoretical work now.

The notions of 'diagnosticity' and informativeness: a Bayesian analysis

There are at least two purposes that an identification parade can serve. It can constitute a reliability check, in the sense that it tests whether a witness is able to identify a suspect arrested by the police. Such an arrest may or may not have been based on a verbal or 'visual description',[192] but this is not material to the purpose of the parade, which is conducted merely to corroborate or disconfirm the hypothesis of identity. Alternatively, the parade may have the purpose of securing an independent piece of evidence against the suspect, in the sense that it increases the perceived likelihood that the suspect is the perpetrator (see Chapter 3 for a related discussion).

Wells and colleagues (Lindsay & Wells, 1980; Wells & Lindsay, 1980; Wells & Turtle, 1986) have provided a Bayesian analysis of parade identifications, although they do not make the explicit

[192] I mean the kind of description inherent in an 'identikit' portrait, or some similar assisted reconstruction of the witness' visual memory of the perpetrator.

distinction referred to above. I will attempt a little later to make clear how this distinction impacts on their analysis.

There are two key notions in their Bayesian analysis of parades. The first is the well known 'likelihood ratio', which has a fundamental place in Bayesian statistics. Lindsay & Wells (1980) refer to this ratio as the 'diagnosticity' of a lineup: how 'diagnostic' the lineup is of the suspect's guilt (or innocence). In terms closer to Bayesian principles, the ratio expresses how much more likely the data are to have occurred given the truth of one hypothesis (that the suspect is the criminal) relative to the other (that the suspect is innocent):

$$\frac{P(ids \mid s = c)}{P(ids \mid s \neq c)}$$

where ids = identification; s = suspect; c = criminal; | is the conditional operator .

An example should make the meaning of the ratio clear. Table 4.3 presents the distribution of witness[193] choices in two sets of identification parades: in one of the parades in each set the suspect is guilty (parades l & n), in the other the suspect is innocent (parades m & o).

	Lineup Member						
	1	2	3	4*	5	6	N/P
l	5	3	6	31	8	3	4
m	8	5	6	5	2	7	27
n	5	3	6	31	8	3	4
o	6	5	6	25	8	3	7

N/P = not present. * = Target

Set {l, m}: diagnosticity = [31/60]/[5/60] = 6.20

Set {n, o}: diagnosticity = [31/60]/[25/60] = 1.24

Table 4.3 Diagnosticity in a series of hypothetical parades.

In the first set of parades, the suspect is 6.2 times more likely to be chosen when she is the criminal than when she is innocent. The identification parade is said to be diagnostic of the suspect's guilt. In the second set of parades, the suspect is only 1.24 times more likely to be chosen when guilty than when innocent, and the parade is thus not very diagnostic of the suspect's guilt.

The absolute size of the diagnosticity ratio is thus the measure of the parade's value, taken independently of all other things that have a bearing on the case.

[193] Note that real witnesses, and not mock witnesses, are the witnesses of interest here.

The theoretical position assumed in the formulation of the ratio is clearly aligned to the second of the purposes I distinguished earlier: a parade provides independent evidence of the guilt or innocence of the suspect. Navon (1990a, 1990b), however, has argued that the diagnosticity ratio is a measure of reliability, and does not provide independent evidence of identity. I will consider his objections in the next section of the chapter.

The second key notion in the Bayesian analysis proferred by Wells and colleagues is that of 'informativeness', or 'information gain'. A positive (or negative) identification presumably leads to some alteration of the probability that the suspect is the criminal from the point of view of the investigating officer, and the amount of this change is the nett informational value of the identification. Information gain is the difference between the prior probability that the suspect is the criminal, and the posterior probability of the same (dependent only on the intervening identification or non-identification of the suspect). It is stated formally as

$$\text{Information gain} = |p(s=c) - p(s=c|ids)|$$

or,

$$\text{Information gain} = |p(s\neq c) - p(s\neq c|nids)|$$

where s = suspect, c = criminal, | = sign for conditional occurrence of an event, ids is the event that an identification is made, and nids is the event that an identification is not made

$$p(s=c|ids) = \frac{p(ids|s=c)p(s=c)}{p(ids|s=c)p(s=c) + p(ids|s\neq c)p(s\neq c)},$$

and s, c, ids are defined as before.

The absolute value of the difference between the probabilities is taken, since it is the size of the gain that is important, and not the direction. The value of the identification depends on the size of the prior probability (although this is only clear from a close examination of the Bayesian ratio): a positive identification might lead to a big increase in the likelihood of the suspect's guilt if it is low to begin with, but will generally have a smaller impact the higher the prior likelihood of guilt.

Whereas the diagnosticity ratio concerns the independent amount of evidence provided by an identification at a parade, information gain expresses the effect that an identification (or non-identification) has on the total amount of information that existed before the identification.

The notions of diagnosticity and information gain attempt to conceptualize the identification task embedded in identification parades in terms of a type of statistical information theory. There is some question about the interpretation of the measures, though. David Navon (1990a, 1990b) has recently argued that there are several problems with these notions, and has suggested a different way of looking at parades in terms of Bayesian theory.

Parade identifications and ecological likelihoods: Navon's objections

In order to understand Navon's objections to the Bayesian formulation provided by Wells and colleagues, it is important to grasp a methodological feature of much identification parade research. Police identification parades are of at least two types: parades where the perpetrator is present, and parades where the perpetrator is absent. In both types of parade, the police have placed a suspect in the parade, but they invariably do not know whether the suspect is guilty. In order to reproduce this aspect of police practice, two identification parades are frequently used in psychological research, one containing the 'guilty' suspect, and one containing an 'innocent' suspect. The difficulty with this procedure lies in the justification of the choice of innocent suspect for the second parade. This is often achieved by choosing a confederate who bears some physical resemblance to the perpetrator, and 'deeming' this foil to be the innocent suspect.

'Fixing the resemblance', in this way, argues Navon, begs the question of the informational value of the lineup. Diagnosticity (which is the ratio of correct identifications in perpetrator-present lineups to incorrect identifications (of the innocent suspect) in perpetrator absent lineups) can be made high by choosing a suspect (and foils) that is highly dissimilar to the perpetrator, and very low by choosing a suspect that is highly similar. Navon contends that similarity is a very important aspect of the identification: it embodies the most useful evidence provided by the identification. The measure developed by Wells and colleagues is only a measure of the reliability of a witness identification - it expresses a ratio of correct identifications to incorrect identifications. The point is that concentrating on the transduction process itself will inevitably mask the nett evidential value of the information, since this depends on ecological likelihoods, and not on the process of transduction. The correct question to focus on is how likely the random match is of the perpetrator and any innocent person whose features happen to match those of the perpetrator. The critical information is not how likely a false alarm is, but how difficult it is to find a foil that will be falsely identified. By analogy, the value of a forensic blood type match depends on how unlikely such a match is, not on how reliable laboratory techniques are for determining the match.

The formulation of diagnosticity given by Wells and colleagues (see page 73) is thus conditional on the resemblance between the suspect and the target, but this conditionality is not taken into account in the formulation. Navon accordingly suggests another likelihood ratio,

$$\frac{p(R_{st} \rangle R_0 \mid S = T \text{ and } N = n)}{p(R_{st} \rangle R_0 \mid S \neq T \text{ and } N = n)},$$

where R_{st} = the resemblance between subject and target; R_0 = the lowest degree of resemblance that can be inferred from the parade, s = suspect, t = target.

Navon makes a few suggestions about obtaining parameters for determining how likely a match is given that the suspect is innocent. In particular, he suggests that the critical practical index is how difficult it is to get a foil who is not rejected by the witness: the target/perpetrator is then at least as close in resemblance to the suspect as he is to the foil who is not rejected. The next step is to find a probability value by retrieving the base rate for the critical foil, namely by comparing him to empirical distributions of face types (although such distributions do not exist), or by counting the number of foil candidates searched until a match is made.

Wells & Luus (1990a) replied to some of Navon's objections, and the replies were countered in a second article by Navon (1990b). The details of the exchange are not important here. What is important is that both parties acknowledge the importance both of ecological likelihoods and of the physical resemblance between the foils and the perpetrator of the crime.

Neither of the Bayesian approaches discussed in this section offer a sufficient conceptualization of the task embodied in the identification parade. In several recent articles, Wells and colleagues (Wells, 1993; Wells, Seelau, Rydell, & Luus, 1994) have attempted a more complete analysis.

Identification parades as recognition tests

Wells et al. (1994) outline a theory of identification parades constituted by two propositions and a corollary. The theory has important consequences for the selection of parade foils; these will be discussed later in the chapter.

> *Proposition 1*: The purpose of a lineup is to uncover information in an eyewitness's recognition memory that was not available in recall.

The identification parade closely resembles a memory recognition test: the witness is presented with an array of choices, and is required to make the correct judgement. Where the perpetrator is present, the correct judgement is to point him out, and where the perpetrator is absent, the correct judgement is to indicate that none of the choices is suitable. A critical difference, though, is that the person who administers the recognition test usually knows the correct answer. In the identification parade, police hope that the witness will lead them there.

Parades are conducted because verbal descriptions of the perpetrator given by witnesses do not contain information that allows us to decide whether the suspect is the culprit or not. This may be due to the poverty of the description, or to its inaccuracy. Several studies have shown that witnesses may provide substantively inaccurate descriptions, but are nevertheless able to identify the perpetrator in a parade (Pigott & Brigham, 1985; Wells, 1985).

Wells et al. directly assert that the function of a parade is to adduce evidence of identification, i.e. to increase or decrease the probability that the suspect is the criminal. The verbal recall given by a witness to the police is a different matter from the test of recognition that the lineup essentially is: the lineup must look for evidence that takes the matter further than the witness's original description, it must provide information about identity in addition to this report.

> *Proposition 2*: The identification process is governed not only by simple memorial factors but also by extramemorial judgement and heuristic processes.

One such heuristic is the so-called 'relative judgement strategy' (Wells, 1984), discussed later in the chapter. A witness may construe her task to be the identification of the person who looks most like the perpetrator, rather than an absolute judgement of identity. This is an efficient strategy when the suspect is present in the parade, but is extremely dangerous when he is not!

Wells et al. provide a slightly broader conceptualization of proposition two by attaching a corollary to it.

> *Corollary to proposition 2*: A lineup task can be likened to a social psychology experiment: Factors that can confound the psychological experiment can confound the lineup task.

This analogy was discussed at some length by Wells & Luus (1990), Wells (1993), and Doob (1980, cited in Wells, 1984). There are obvious parallels between the need for 'blind' and 'double blind' procedures in both cases, the need to take random sampling variation into account, and so on. I will not detail the analogy any further here, except to note that the explicit comparison with social psychology is too specific - experimental methodology in general is probably a more suitable analogue.

On the basis of the two propositions and the corollary, Wells et al. make a number of recommendations for the conduct of identification parades.[194] These are summarised below as Table 4.4.

[194] But see Wells (1988) for a similar set of recommendation without the theoretical basis.

	Recommendation	Justification[195]
1	Verbal descriptions should be taken from all eyewitnesses.	To determine whether a parade has taken the issue of identity further than the witness's original description
2	For every suspect there should be five distractors in the parade.	Arbitrary floor on probability of correct random witness choice (in analogy to a type I error rate).
3	Distractors should be chosen to match the witness's description of the suspect.	The parade attempts to gain more information than present in the original description.
4	Separate lineups should be used for each eyewitness in multiple-witness crimes	Independence of identifications (faults in parade structure will be replicated if unchanged).
5	The positioning of the suspect should be different for each lineup in which he appears	In analogy to counterbalancing procedures.
6	Different distractors should perhaps be used where a suspect appears in more than one parade	In analogy to counterbalancing procedures.
7	Lineup administrator should not know who the suspect is, nor who the distractors are.	In analogy to double-blind procedures.
8	Eyewitness to be given the 'if present' instruction.	To make the response criterion stricter.
9	The witness should be asked in two stages; viz. i) can you identify the perpetrator?; ii) who is it?	To make the response criterion stricter.

Table derived from Wells, Seelau, Rydell & Luus (1994).

Table 4.4 Recommendations made by Wells et al. for the conduct of identification parades

The theory and recommendations set out immediately above assumes that the identification parade is intended to provide independent evidence of identity, and sets out ways to optimize the reliability of parade procedure. It fails to consider the alternate conceptualization of the parade as a test of witness reliability. I have previously referred to the distinction between the conceptualization of the identification parade as a test of reliability, and as a method of acquiring independent evidence of identity. This distinction is important, because it has different implications for the evaluation of parade identifications.

Navon (1992) has recently made a similar distinction, and he notes some of the implications that each conceptualization has for parade practice and evaluation. The distinction he draws is between the following purposes of a parade: i) to provide an identification of the offender; or ii) to provide information about the resemblance between suspect and perpetrator beyond that provided by the fit with the original description. The first of the purposes is akin to a test of reliability, since it aims at proof of identity - an identification confirms that police have 'the right man'. Navon refers to this as the 'high match approach' (p. 584). The second of the purposes is intended to gather information

[195] Some of the justifications in this table do not appear in the article by Wells et al., but nevertheless flow naturally from the propositions and corollary.

from the eyewitness, and does not intend to secure an identification. This is the so-called 'minimalistic approach'.

The two purposes have very different implications for the selection of parade distractors. In this respect, Navon shares the idea with Luus & Wells (1991, Wells et al. 1993) that distractors should be chosen to resemble the original description of the offender, and not the suspect. Since this view is very different from much existing police practice and psychological research, I will introduce it after discussion of earlier work on the selection of distractors for parades.

Research on distractor selection and evaluation

In Chapter 3, I outlined a central legal requirement for the construction of identification parades, namely that the suspect be placed alongside other people of approximately the same height, weight and physical appearance. This requirement is long established in both English and South African law. It is a difficult criterion to evaluate, and some research by Malpass and colleagues (Malpass, 1981; Malpass and Devine, 1983) has attempted to apply the principles used in measures of lineup fairness to the problem.

Malpass & Devine (1983) suggest a method for evaluating the suitability of individual foils, which is closely related in principle to the measure of effective size. The critical datum for evaluating the suitability of an individual foil, Malpass & Devine suggest, is the extent to which the foil is chosen below chance expectation in a mock witness task. Thus, if foil 1 is chosen from a ten member lineup with some low probability, this would suggest that the foil does not adequately represent the description given to the police by the eyewitness. (The question of the extent of departure from chance expectation is a thorny one. Malpass & Devine suggest leaving the decision to the fact finders).

An alternative approach to measuring parade fairness would then be to set a minimum size criterion, and to determine whether the parade meets the minimum size (the estimate of size would be determined by including only plausible foils - and the suspect - in the total). This approach would apparently be intuitively appealing to lawyers and judges (Malpass and Devine, 1983). I argue in Chapter 6 that decisions about whether an individual foil meets some minimum identification criterion should be made in terms of a suitable probability model.

The point to note here is that although this method of evaluating individual foils does not presuppose a particular strategy for choosing foils, psychological research using the method has followed the legal model of matching the foils to the suspect. Wells, Navon and colleagues are adamant that this approach is not only inoptimal, but courts grave danger. In order to present the argument clearly, I

discuss the traditional method of foil selection, and its suggested replacement, under separate headings. It is not necessary to distinguish the approaches taken by Wells and Navon, since they overlap.

The match to suspect strategy

This is the method advised by the courts. A number of foils are selected to serve as members of the parade, on the basis of their physical similarity to the suspect. A number of protections appear to be intended (Luus & Wells, 1991): in the first place, the probability of a correct, random choice is reduced; secondly foils allow the identification of known error, which reveals that the witness's memory is faulty; and thirdly, properly chosen foils assure that eyewitnesses cannot use deductive reasoning to determine which member of the lineup is suspected by the police, i.e. it reduces suggestion.

There are several problems with this approach. The most pressing is that there is no clear optimal 'point of physical resemblance' which can be used as a selection heuristic. On the one hand, a low degree of physical similarity will allow the witness to identify an innocent suspect on superficial features the suspect shares with the offender.[196] On the other hand, too great a degree of physical similarity will render an identification impossible, even if the witness has an accurate memory of the offender. Wells and Luus (1991) offer a *reductio ad absurdum* here, namely the situation where the parade consists of a suspect and five of her clones.

The match to suspect strategy is intended to protect against situations where the suspect happens to resemble the offender more than the foils do. It is not clear that it provides this protection. It might be possible to find foils that bear considerable resemblance to the suspect, but it will be impossible to guarantee that the foils share all the features that an innocent suspect shares with the offender. Since the police are (partially) likely to arrest the suspect on the basis of some resemblance to the original description, and since it is not possible (or desirable) to completely match the foils to the suspect, the suspect is likely to resemble the offender more than any of the foils. An innocent suspect is therefore not completely protected against a superficial resemblance with the offender, although the dangers attendant upon such resemblance will be somewhat reduced.

[196] See the cases of *Pelwan* and *Masemang*, which were discussed in Chapter 3.

The match to description strategy

A better approach to distractor selection, argue Wells, Luus, Navon, and colleagues, is to match them to the description of the offender originally given by the witness. This is in line with the notion that one of the primary functions of an identification parade is to adduce evidence of identity over and above what the witness is able to recall, i.e. the parade serves as a test of recognition memory.

The major advantage of the match to description strategy is that it apparently solves the problem of suspect - foil similarity. It protects against superficial resemblance of suspect and offender by ensuring that the foils share all the features of the offender that the witness recalls. It specifies the physical features to be shared by all lineup members. On the other hand, it also avoids the problem of excessive suspect - foil similarity: since the original description cannot be a complete depiction of the offender, foils and suspects will differ on the many features not specified in the description. It specifies the physical features not to be shared by parade members. In short, the match to description strategy ensures that there is adequate 'feature heterogeneity' (Gibson, 1969), or 'propitious heterogeneity' (Wells, 1993).

The guarantee of 'propitious heterogeneity' is postulated as the greatest advantage of the match to description strategy. In particular, Luus & Wells suggest that where parades formed under either strategy will reduce false identifications of innocent suspects, the match to description strategy will increase accurate identifications of guilty suspects. A recent paper by Wells, Rydell & Seelau (1993) reports a set of experiments that appears to bear this contention out, but it is worth waiting for replications before making much of the comparison.

Several objections can, in turn, be lodged against the match to description strategy. These, and the answers Luus & Wells (1991) provide, are worth detailing here, since one of the central aims of the thesis is to develop a measure of facial similarity for use in assessing the fairness of identification parades, and the application of the measure will adopt a 'match-to-suspect' strategy.

The first problem with the strategy concerns situations where the police arrest the suspect on grounds other than the initial witness description, and conduct a parade in which the suspect is very dissimilar to the description. In this type of situation, the match to description strategy is likely to result in a parade where the suspect is distinctive, and the foils fairly similar - both to each other and to the description. Luus & Wells suggest that the two selection strategies be used in conjunction here: i.e. that foils are chosen to share features with both the suspect and the original description.

The second problem concerns situations where the original description is idiosyncratic (e.g. where it mentions a tattoo). Luus & Wells suggest the use of artwork or a concealing patch, or alternatively,

not conducting a lineup, since the description is sufficiently selective to suggest that the suspect and perpetrator are one and the same person.

The third problem concerns cases where there are multiple witnesses and each provides a different description. Here Luus & Wells suggest using separate lineups, although it is not clear how this can be satisfactorily achieved when there is only one suspect!

Finally, there is the problem where some feature of the suspect which the witness did not recall of the perpetrator is very distinctive, and may lead of its own to the identification of the suspect. Luus & Wells point out that each foil is likely to have some distinctive feature anyway, and this will counter the suspect's distinctive feature.

The match to description strategy coheres in a satisfying way with the theoretical conceptualization of the parade as a recognition test. However, it requires auxiliary strategies, and it is not clear that its advantages over the match to suspect approach are substantial. There is empirical evidence from a study by Wells et al. (1994), which suggests that the match-to-description strategy results in more accurate identification decisions, but this unreplicated study is the only evidence of this kind to date.

Research on structural aspects of identification parades

Psychologists have often pointed out the similarity of the identification parade to a rudimentary scientific experiment: the police have an hypothesis about the identity of the perpetrator, and they follow a procedure which tests this hypothesis. The procedure itself - i.e. the parade - has some of the characteristics of experimental control in so far as it attempts to rule out several sources of confound (Wells & Luus, 1990; Wells, 1993; Wells et al. 1994). Nevertheless, it is perfectly legitimate to question whether the control is rigorous enough, and indeed, whether present procedures are structurally optimal. The analogy of lineup and experiment is so arresting that psychologists have directed much of their effort here.

There is much to discuss; I will present a somewhat simplified account, progressing from research of less consequence, to some that has profound implications for the conduct of identification parades. I begin by showing that some of the protections already built into identification parade procedure are well placed, and that the absence of these protections (and therefore the failure to ensure that they are in place) would certainly lead to a higher rate of mistaken identification decisions. In particular, instructions to witnesses indicating the possible absence of the criminal are important, and the courts are also quite correct in approaching voice identifications with a considerable degree of skepticism. I also discuss empirical results from studies that examine alternative types of identification parade, and which show that present structures are inoptimal and could be gainfully modified.

Parade Instructions

In a famous set of studies, Orne (1962), and Rosenthal (1966) argued (and showed) that human subjects are motivated to determine experimental hypotheses, and often do so by attending to very subtle cues inherent in the nature and structure of the experimental situation. Humans are not the inert objects postulated by certain traditions of scientific enquiry, but are indeed, highly reactive. Buckhout (1974), and Wells & Luus (1990) have argued that identification parades must suffer from the same problem. Witnesses understand why they have been brought to the parade, and they may well assume that it is their responsibility to point someone out. They may construe their task to be something like 'find the parade member whom the police suspect', rather than 'is the person who committed the crime present?' Subtle cues from police officers, or biased instructions - neither of which need be conscious attempts by the police to subvert the fairness of the test - among other things, may affect the witness's choice.

Of course, these threats are well known and carefully protected against in most legal systems. South Africa's record in this regard is, in particular, good. Nevertheless, it is interesting to note that empirical studies have explored the extent of this threat, and thus, several staged-crime experiments have shown that instructions that presuppose the presence of the criminal in the parade ("Point out the person who committed the offence") lead to much higher error rates than instructions that don't ("Note that the criminal may not be present. If the criminal is present, point him out") (Malpass & Devine, 1981a; Cutler, Penrod & Martens, 1987a; Cutler & Penrod, 1988). The implication is that the instructions given to witnesses ought to be tempered quite carefully with a warning indicating the perpetrator's possible absence,[197] which is an approach favoured by South African courts. Paley & Geiselman (1989) compared a set of clearly biased instructions with those usually given by the LAPD (Los Angeles Police Department), and with a further set of 'more balanced' instructions, and found an increase in mistaken identifications for the biased instructions, but no such increase with either of the other instructions.

There are several other threats to the fair conduct of an identification parade that inhere in the social nature of the task. Malpass & Devine (1984) cast the performance of a witness at an identification parade in terms of subjective expected utility: the witness's decision to identify someone will depend on more than just the quality of his memory, but indeed on a number of considerations extraneous to the act of witnessing the event. Ainsworth & King (1988), for example, interviewed a number of witnesses who had recently attended identification parades, and reported an extremely high degree of

[197] However, one study suggests that this effect is dependent on whether subjects are 'debriefed' prior to the identification or not. Subjects who have not been debriefed fail to show the observed effect, and seem in general to take the task much more seriously than undebriefed subjects (Köhnken & Maass, 1988).

fear of reprisals. This fear is no doubt exaggerated by the requirement in both the English and South African legal systems that witnesses touch the person they wish to identify.

Malpass and Devine argue that the events eyewitnesses provide testimony of are often viewed under inoptimal conditions, and that witnesses typically face some degree of ambiguity when deciding whether to make an identification. Witnesses must take some risks whatever they decide. Subjective expected utility theory postulates that a decision to identify will be the result of a process that weights each available alternative (i.e. to choose, or not to choose) according to the perceived value of the outcome; it is quite possible therefore that a witness with a faultless memory of the perpetrator may refuse an identification, or that a witness who remembers nothing of the perpetrator's identity may attempt an identification.

There is little of practical value for the police or the courts in this observation, since it is recognized in several ways by these bodies: however, it is worth reminding ourselves that the quality of the witness's memory is not the court's only concern in evaluating eyewitness identification evidence.

Intervening mugshots

In Chapter 3 I noted the concern which the courts and legal commentators have expressed regarding the practice of showing witnesses photographic 'mugshots'. The concern is that mugshots intervening between the witnessed event and an identification parade may 'contaminate' a witness's memory and lead the witness to mistake her memory of the mugshot for her memory of the perpetrator of the crime.

Several studies have explored this possibility, but the results are unclear regarding the potential for 'contamination'. Studies by Loftus & Greene (1980), Jenkins & Davies (1985), and Brigham & Cairns (1989) suggest that such contamination is likely to occur, but studies by Gorenstein & Ellsworth (1980) and Lindsay, Nosworthy, Martin & Martynuck (1994) do not support this conclusion.

Gorenstein & Ellsworth argue that any observed effects are likely to occur by one of two routes. In the first, a witness will incorrectly identify a suspect because his face has become familiar as a result of the mugshot interpolation, but the witness cannot place the situation in which he first observed the face. In the second, a witness will feel committed to a choice made during the mugshot interpolation, and will therefore choose the suspect again at the identification parade.

Of the studies which have tested the effect of interpolated mugshots, that by Lindsay et al. is probably closest to actual police practice: a large set of photographs was used, and effects were tested for several confederate suspects. That they failed to find any substantive adverse effect is noteworthy.

Lindsay et al. also outline a novel way of using mugshot albums, which has direct bearing on some contentions discussed earlier in the chapter. Instead of attempting to secure a positive identification from a mugshot album, Lindsay et al. suggest that witnesses be asked to identify all the mugshots they cannot certainly exclude. This will provide police with evidence concerning the physical identity of the perpetrator, but not with an identification. Such evidence should not be admitted in trials, but used by police to further investigate the case at hand. In several tests of this strategy, Lindsay et al. claim to have shown that the accuracy rate is greatly increased: the confederate perpetrator is almost always included in the set of mugshots that witnesses are unable to reject, and this set is also usually manageably small.

This is a very interesting recommendation, and it has a direct parallel with the minimalistic approach to identification parades favoured by Navon, as discussed above. It seems to suggest that the similarity of mugshots (parade foils) is crucial, but that this similarity works in a different way to that imagined by the courts. In particular, similarity of mugshots (parade foils) hinders the accurate identification of the perpetrator. This must be a matter of degree, though, since extreme dissimilarity of parade members self evidently prejudices the suspect. I suggest that we may need to revisit the notion of an 'optimal similarity function' rejected by Luus & Wells (1991), and I will directly investigate this possibility in a later chapter.

Size of the lineup

Most legal systems make explicit recommendations regarding the minimum number of persons required to constitute an identification parade. In England and South Africa, this number is frequently said to be eight (see Chapter 3). No explicit justification is given for this number, but it presumably protects the suspect against random or haphazard identifications.

Psychological research has in several instances explored the effect of different parade sizes on measures of witness accuracy. Thus, Nosworthy & Lindsay (1990) systematically varied the size of a parade by adding either poor foils or good foils, i.e. nominal size was increased, but effective size held constant, or both nominal and effective sizes were increased. They found that neither manipulation affected either i) accurate identifications of the perpetrator, or ii) mistaken identifications of innocent suspects. They concluded that a parade consisting of the suspect and only three foils may be adequate, in the sense that it does not affect the ability of the witness to identify the perpetrator when he is present, or the ability to reject the parade when the perpetrator is not present.

This line of argument was taken even further by Gonzalez, Ellsworth & Pembroke (1993), who showed in a set of experiments that 'showups' i.e. a parade consisting of only the suspect, did not

exhibit an increase in mistaken identifications relative to properly formed parades, nor did they affect the rate of accurate identifications. If anything, performance of witnesses at showups was characterized by fewer mistaken identifications! They explain this anomalous finding in terms of an 'absolute judgement strategy', a notion more fully elaborated by Wells (1984), and discussed later in the chapter.

Wagenaar & Veefkind (1992), on the other hand, also conducted a study in which many-person parades were compared to one-person parades, and they found a substantially increased rate of false alarms in the latter case. They warn against the practice of showups, declaring it unsafe.

The studies by Nosworthy & Lindsay, and Gonzalez et al. report interesting results, but they appear to overlook the preventive function of identification parades. Parades with very few members cannot protect against witnesses who make random or haphazard identifications, even if such witnesses are extremely rare. Parades of considerable size may not maximize the rate of accurate identifications relative to smaller parades, but they better prevent the dangers created by the kind of witness who feels that he must make an identification, even if such a witness is more likely to attempt an identification when confronted with a parade than a showup. In the former case, mistakes can be diagnosed (i.e. when the witness identifies an innocent foil), but in the latter case, an identification cannot be diagnosed, and may have severe consequences for the innocent suspect.

Modes of presentation

The identification parade is usually a test of visual memory, but most legal systems also recognize that cues to identity may reside in the voice, in mannerisms, and in gait. Explicit recognition is given along these lines in South African law, in, among others, the cases of *R. v. Gericke*, *S. v. M.*, and *R. v. Chitate*. (See Chapter 3).

The case of voice identification has attracted the most attention in the psychological literature, and comprehensive summaries can be found in Bull and Clifford (1984), and Yarmey (1994). The most pertinent finding here has been that voice identifications are prone to yield alarmingly high false alarm rates: voices are especially easy to confuse. It is difficult to place anything resembling a precise estimate on how probable a false alarm is in a voice parade, since the experimental conditions employed in different experiments differ greatly, as do the false alarm rates themselves. This is a necessary difficulty: in order to determine which structural factors are critical to the rates of accurate and false identification in identification parades, conditions are purposefully manipulated in distinct ways. However, an experiment by Melara, De Wit-Rickards and O'Brien (1989) provides a direct comparison of voice identification parades and visual identification parades, which is illustrative of

the point: visual parades had an associated false alarm rate of 0.44 ('hit rate' = 0.29), while voice parades had an associated rate of 0.85 ('hit rate' = 0.6). (Both parades were preceded by the same staged crime, making the error rates comparable). Similar differences between visual and voice parades are reported by McAllister, Dale & Keay (1993). McAllister et al. also report that subject jurors are not intuitively aware of the differences in accuracy between eye- and ear- witnesses, and are not easily persuaded of this, either.

Identification parades typically used in Western countries restrict the presentation of the parade to one mode, usually the visual. This arrangement is usually out of keeping with the nature of the witnessed event: a witnessed event will provide many cues to identity over and above those afforded by the static view re-created in a visual identification parade.[198] In the second place, cues from a single (sensory) domain may be degraded with respect to the original context (e.g. physical disguise), and it will be easier to identify the perpetrator with the assistance of cues from other domains. If we accept the truth of these propositions, and modify parade structure, what benefits will multi-mode parades have for witness ability?

Two recent sets of experiments, reported by Melara et al. (1989), have explored the benefits that accrue to parades from modifications to the structure of the task that enable the joint presentation of cues from different domains (e.g. voice, gait, clothing). These modified parades make for fairly substantial reductions in the false alarm rate. Melara et al., for example, showed that lineups which allowed witnesses to view and hear parade members reduced false alarm rates from 64% to 38%, and from 54% to 36%, in two separate conditions.[199]

The idea of allowing witnesses access to cues that are not solely visual in nature is already in practice in Sweden. The parade is formed in a room in which parade members are allowed to sit down, to smoke, to communicate, and to behave in general as they ordinarily would over an extended period of time. The parade is viewed by the witness from behind a one-way mirror, and members of the parade are not aware of when the witness is brought in to view the parade (Shepherd et al., 1982).

Thus far I have considered the question of different sensory modes, but it is also important to note that parades presently differ within single sensory modes. What I mean is the practice of eliciting identifications from photographic parades as an alternative to corporeal lineups. In some parts of the

[198] Some associationist theories hold that memories are encoded in a network of associations, and the more associations available to a witness, the greater the likelihood that the attempt to remember will be successful. See, for example, Anderson & Bower (1973). There is also evidence to suggest that cross-modality information is particularly powerfully encoded. Constraining the domain of information must therefore impede the ability to remember.

[199] Note however that this reduction in false alarm rates in dual mode lineups was not present when the target was absent from the lineup.

United States, for example, photographic parades are favoured over corporeal parades (Goodman, personal communication).

Intuitively, corporeal parades would seem to have an advantage: the detail to be obtained from an *in vivo* inspection of the suspect must surely far exceed that present in a photograph. However, the consensus results from several studies indicate that this is not the case: that there is either very little difference between the capacity of photographic and corporeal parades to elicit identifications, or that there is no difference at all.

Thus, Shepherd, Ellis and Davies (1982) staged a crime in front of 242 subjects, and compared their identification accuracy rates in four different parade modes: live, video, black and white photographic stills, and colour stills. No significant differences were observed between accuracy rates obtained from these different parades: indeed, the results slightly favoured the video parade. Similar results are reported by Dent (1977), among others.[200]

Cutler, Berman, Penrod & Fisher (1994) have recently broadened the conceptualization of identification methods, subsuming the many methods of securing identifications under the term 'identification test media'. They report an extensive quantitative review of the literature on 'identification test media', and draw the conclusion that there is no notable difference in the literature between live, video, or photo- lineups. There is also no evidence to suggest that media which embellish cues[201] aid identification accuracy or reduce false alarms.

Multiple suspect lineups

The characteristic identification parade, we imagine, consists of a single suspect and a number of innocent distractors. In practice this is not the case, as many trial lawyers will tell you: more often than not a lineup contains multiple suspects. In one Cape Division case, for instance, an advocate represented three of fifteen accused, all of whom appeared in a 43-man parade.[202]

It is evident that a parade containing several suspects has attendant threats to its validity, perhaps of a different form to those associated with a single suspect parade. Wells and Turtle have investigated these threats, and compared the value of the information yielded by multiple suspect parades to that

---

[200] But for a dissenting opinion see Egan, Pittner, & Goldstein. (1977).

[201] Several studies have explicitly manipulated the nature of cues present in identification test media. In some cases cues have been embellished - Cutler & Penrod (1988) allowed witnesses to view the gait of video parade members - and in others, cues have been degraded. Cue degradation has typically been effected by disguising the suspect, e.g. Cutler, Penrod & Martens (1987).

[202] The practice of using multiple suspect lineups is not uniquely South African. In fact, Wells reports that in his visits to several dozen police precincts in the United States in the 1980's he found this to be more often the case than not (Wells & Turtle, 1986).

yielded by single suspect models, within the Bayesian statistical framework set out earlier in the chapter.

At the heart of it, the difference between the two parade models involves a trade-off between the increased likelihood of achieving an accurate identification in the multiple suspect parade, given the presence of a greater number of suspects,[203] and the increased likelihood of a false identification in the same parade. In the case of a single suspect parade, any identification of a parade member other than the suspect is known to be false, whereas in a multiple suspect parade the witness can make a number of undetectable mistaken identifications. Specifically, Wells and Turtle have shown that only when the population of potential suspects is very small, is the multiple suspect model to be preferred, otherwise it poses a greatly increased risk of mistaken identification. In practical terms, such a situation is unlikely - i.e. the situation where the police could, for example, say "One of five people did it" - and a consideration of the risks would seem more pertinent in the evaluation of the multiple suspect parade.

There is another risk associated with the multiple suspect model, often explicitly overlooked by courts in South Africa, and elsewhere. One of the underlying features of an identification parade is the protection it gives suspects against random, or haphazard, identifications. This protection is built into the foil-suspect ratio considered acceptable by courts, which in South Africa is often stated as eight to one. If additional suspects are added to the parade without preserving this ratio, it follows that the likelihood that a suspect will be identified by random guessing will be increased. Yet, courts do not insist on preserving this ratio: in England, a 1969 circular to the Home office, and in South Africa the law - on the ground, at least[204] - says that it is not necessary to preserve this ratio. There is little to suggest that identification parades invite haphazard identifications, but the possibility of this occurring occasionally must be strong, and the protection provided by the suspect-foil ratio is therefore important.

The principal reason for the use of the multiple suspect lineup by police is probably the efficiency of such a parade. As I noted earlier, the construction of a parade is an exacting and onerous task; so much more if several parades must be constructed for one investigation. Nevertheless, I suggest that the benefits of this type of parade are few in relation to the costs, and everyone would probably be better off without it.

---

[203] I restrict myself - as do Wells and Turtle - to the case of one perpetrator and multiple suspects. The case of multiple perpetrators and multiple suspects is essentially similar, if somewhat more complicated.

[204] This has been pointed out to me by several members of the Cape Bar.

The value of non-identifications

In Chapter 3, I noted the approach of South African courts to the evidential value of non-identifications. Specifically, courts have stated that the failure of witnesses who could be expected to identify the perpetrator must be evidence in favour of the accused - it must imply some doubt as to the evidence that the accused is the perpetrator.

Wells & Lindsay (1980) have investigated the value of non-identifications in terms of their Bayesian formulation of the 'informativeness' of parade identifications, and they come to the same conclusion. However, they go further, and specify the nature of the relationship between identifications and non-identifications: identifications are more informative than non-identifications when they are rarer, and vice versa.

Although courts have noted the significance of non-identifications, it is not certain that they will always accord such failures the appropriate weighting. Wells & Lindsay note the well known tendency of human subjects to overlook disconfirming instances, the similar and widespread inability to accurately interpret base rates, and several other findings from the decision making literature that would lead one to doubt that cognisance at the level of case law is sufficient.

McAllister & Bregman (1986, 1989) and Leippe (1985) have specifically investigated the role that information about non-identifications plays in jury[205] decision making in simulated crime scenarios. The results from these experiments are unfortunately inconclusive, but do not seem to suggest a tendency of jurors to under-utilise non-identifications from eyewitnesses.

Similarity

At the centre of the legal conceptualization of the identification parade is the notion of physical similarity, and the assertion that a parade should be formed so that the foils and suspect sufficiently resemble each other. In this section of the review I discuss psychological research that has directly addressed the issue of similarity. I intend this as preparation for a lengthier section in Chapter 5, where I will attempt to set out a particular approach to the measurement of facial similarity.

One of the clearest demonstrations of the importance of physical similarity is to be found in a study reported by Malpass & Devine (1983). The study uses the mock-witness method detailed earlier in the chapter.

---

[205] Juries were abolished in South Africa more than 25 years ago. The interpretation of jury research to South African legal practice is therefore quite complicated, but I will nevertheless refer to such research at several places in the thesis.

Full length photographs of college age white adult males were rated by twenty judges on six dimensions, each dimension consisting of three levels. Subjects were asked to choose the descriptors which best characterized the photographed person. Table 4.5 presents the dimensions and levels judges were required to use.

Dimension	Level 1	Level 2	Level 3
Hair colour	blonde	brown	black
Hair length	short	medium	long
Hair style	straight	wavy	curly
Height	short	medium	tall
Build	thin	medium	husky
Eye colour	light	dark	-

From Malpass and Devine, 1983: p 89.

Table 4.5 Dimensions and levels used by judges in Malpass & Devine's (1983) study

Each photograph (parade member) was given a total score, per subject, by weighting each level by the ordinal number associated with it within the levels (i.e. level 1 descriptors were weighted as 1, level 2 descriptors as 2, etc.), and by summing the descriptor weights. An average was taken for each photograph, across subjects. In this way, each photograph was assigned a score ranging from 6 to 18. The perpetrator in this experiment matched the six level 2 descriptors, and difference scores were calculated between descriptor total scores and the perpetrator total of twelve. These difference scores were used as suspect-foil similarity measures.

Four identification parades were then constructed so as to possess varying degrees of physical resemblance, and mock witnesses were asked to identify the suspect (on the basis of a description of the suspect, only). The results of the experiment are reported in Table 4.6, below:

Parade	Suspect - foil discrep.	No. of mock witnesses	Suspect	Foil 1	Foil 2	Foil 3	Foil 4	Foil 5
A	8.45	67	0.12	0.37	0.36	0.06	0.05	0.05
B	10.06	66	0.20	0.57	0.08	0.08	0.06	0.02
C	16.98	68	0.80	0.12	0.04	0.03	0.02	0.00
D	23.78	39	0.85	0.03	0.03	0.10	0.00	0.00

Expected proportion = 0.167. From Malpass & Devine, 1983, p 93.

Table 4.6 Distribution of (proportionate) mock witness choices as a function of suspect-foil similarity.

Visual scrutiny of the results of this experiment shows that physical similarity is strongly related to the frequency with which suspects are chosen. Fairness depends on the degree of suspect-foil similarity - indeed, the correlation between suspect-foil dissimilarity and suspect identification probability is perfect[206] - and even small variations in similarity appear to affect the distribution of choices.

Notice especially one feature of the distribution of mock witness choices, which is that lineups A and B contain foils who are identified more frequently than the suspect himself. This anomalous result may have something to do with the nature of the 'mock witness' task - since mock witnesses have not seen the perpetrator of the event, but instead are required to guess the identity of the perpetrator from a verbal description, it may be that the foils in question are more typical of the description than the suspect.[207] The anomalous result in Malpass & Devine's study is not unique. At least two other studies report identification parades in which foils are chosen more frequently than the suspect (Wells and Lindsay, 1980; Buckhout, Rabinowitz, Alfonso, Kanellis and Anderson, 1988[208]). These studies underscore the crucial importance of physical resemblance in identification parades. In particular, they suggest that physical resemblance shouldn't be considered simply in terms of the relative similarity of parade members, but also in terms of the distinctiveness of parade members in relation to a population norm.

Other studies have found results analogous to those reported by Malpass and Devine. Lindsay & Wells (1980) report a staged crime experiment in which they explored the relationship of lineup similarity to Bayesian measures[209] of the information value of identifications obtained from lineups.

The perpetrator of the 'crime' staged for the experiment was a Caucasian in his 20's. He had light brown hair and a moustache. A 'high foil similarity' lineup was constructed by filling it with foils who were "about twenty years old with brown to blonde hair and moustaches" (p 307); and a 'low foil similarity' lineup was constructed by filling it with foils who were "in their late 20s with full black beards and black hair" (p 307). In addition, this second lineup consisted of two Orientals and three Caucasians (apart from the suspect). The success of the similarity manipulation was checked by using a mock witness task - subjects were required to guess the identity of the suspect from a very brief and

---

[206] Note that this correlation is inflated by the fact that the suspect-foil similarity measure is an averaged value.

[207] Imagine that a suspect is described as tall and swarthy. This characterization will fit many people, and it will be more accurate for some than it is for others. The claim I make here is that it is possible that a description may well be true of the suspect, but a better characterization of one or more of the foils.

[208] The anomaly is particularly interesting in Buckhout et al.'s study, since the identification parade that produced the anomalous result was one conducted by police and submitted as evidence to a New York court, and not a parade conducted as part of a psychological experiment.

[209] In particular, measures of 'diagnosticity' and 'information gain'. These are discussed earlier in this chapter (see page 72 onwards).

general description of the perpetrator. Subjects were able to identify the suspect with comparative ease in the low similarity lineup, but not in the high similarity lineup.

More recently, Wagenaar & Veefkind (1992) investigated the effect of similarity on identification performance in both 'target present' and 'target absent' parades. Here, as in most instances of parade research where similarity is incorporated in the design, suspect - foil similarity was assessed by independent judges on a Likert-type scale. Wagenaar & Veefkind report that similarity systematically decreased recognition accuracy: in target present parades, increased similarity led to greater choosing of foils, and in target absent parades, it led to greater false alarm rates. However, the authors do not present complete results for the experiment, and it is not possible to tell whether increased similarity led to fewer correct identifications of the target in target present lineups.

The studies by Lindsay & Wells (1980), Malpass & Devine (1983), and Wagenaar & Veefkind (1992) are not the only studies that have manipulated the physical similarity of foils and suspects. Indeed, similarity is often used as an independent variable in identification parade research, but usually in combination with other variables. Few studies have directly addressed the relationship of foil - suspect similarity to recognition accuracy. In addition, almost no studies have examined the impact of 'facial distinctiveness' on identification performance, even though a large number of studies in the related area of face recognition and perception have shown that measures of distinctiveness are strongly related to recognition ability. I will return to this point later in the thesis, when I review the face recognition literature in search of methods of measuring similarity.

The point to be made in summary is that the psychological literature provides us with clear evidence of the influence of physical similarity on the fairness of the parade: it is critical, and the imprecise way in which it is made a requirement for the conduct of identification parades bodes poorly for legal practice.

Blank parades

One of the most enduring of all the findings in research into identification parades is the relatively low observed rate of accurate identifications and the accompanying relatively high observed rate of false alarms. Thus, in a meta-analysis of some 50 studies using identification parades, Shapiro & Penrod (1986) report a cumulated average accuracy rate of 0.52. Cumulated false alarm rates are not reported by Penrod etc., but in published studies they are rarely below 0.25. This implies that witnesses are rarely more than four times as likely to identify a guilty suspect as an innocent one, and usually only twice as likely to do so.

This latter statistic presents identification parades in a rather bleak light. What is responsible for the poor performance of witnesses? Psychologists have attempted to explain the high false alarm rate in terms of a 'relative judgement strategy' (Wells, 1984), and several structural manipulations that substantially attenuate the false alarm rate support this notion.

Wells (1984) coined the term 'relative judgement strategy',[210] specifically suggesting that the structure of a police identification parade invites a relative judgement on the part of witnesses. The witness knows in the first place that the police would not conduct a parade if the police did not have a suspect firmly in mind, and thus interprets the task as requiring him to identify the parade member that the police suspect. Worse still, the identification task in its present form explicitly involves comparing the members of the parade. This will lead to a decision that is based not on the virtual correspondence of the witness's memory of the suspect and the parade members, but on the relative correspondence of the witness's memory and the parade members. The task will in short be interpreted as 'finding the parade member who, among those present, looks most like the perpetrator of the crime', which is not the task it is intended to be.

Wells argues that this is almost certainly a simplified portrayal of a witness's choosing behaviour at an identification parade - witness's choice strategies are likely to vary along a dimension characterized at the extremes by relative and absolute judgements - but he insists that it also identifies a very real problem.

One solution to the problem of relative judgements of physical similarity is to precede an identification parade with a 'blank parade', in which all the parade members are known to be innocent. Witnesses who choose someone from this lineup are clearly mistaken, and to be dismissed as too unreliable to complete the main identification task.[211] Wells successfully employed a blank parade in a staged crime experiment,[212] and discovered that subjects who chose a member of the blank parade were almost twice as likely to make an incorrect identification from the subsequent lineup, and likewise were only about half as likely to make a correct identification from the subsequent lineup.

Wells argues that the results of this experiment show that identifications from lineups are adversely affected by relative judgement strategies. Wells points out that blank parades are not feasible in

---

[210] Wells coined the phrase, but the idea is not originally his. Glanville Williams, the famous English legal authority, expressed a similar concern in a 1963 paper:

> The witness may ... be inclined to pick out someone, and that someone will be the one member of the parade who comes closest to his own recollection of the criminal. (p 487).

Woocher (1977) also made several remarks which have much the same meaning as that intended by Williams.

[211] This particular solution is also favoured by Glanville Williams (Williams and Hammelmann, 1955, p 487).

[212] Note that Wells exposed all subjects to the second lineup, whether or not they made accurate identifications in the first lineup.

practice. Witnesses would quickly learn that an identification parade consists of two parts, one involving the real suspect, and the other not, thus rendering the blank parade ineffectual. Also, if police find one parade difficult to construct - as they certainly do - then constructing two parades would be doubly difficult. The solution must lie elsewhere, in some other manipulation that counteracts the relative judgement strategy.

It is interesting to note, in passing, that Lindsay & Wells (1980) suggest that the protection afforded to suspects by increased similarity of foils is due to the counteraction of the relative judgement strategy. The physical features of the criminal represent the source of the original memorial representation, and will draw witness choices at a fairly high rate. When the parade is properly constructed, an innocent suspect will not resemble the perpetrator more than the foils, so the relative judgement strategy will not lead to identification of the suspect. It follows that when the innocent suspect and foils are dissimilar, and the suspect resembles the perpetrator (which is probable), the relative judgement strategy will lead to identification of the suspect.[213]

Sequential parades

A second, and highly successful structural manipulation, introduced by Lindsay and Wells (1985), is the so-called 'sequential parade'. This alteration to the lineup task attempts to counter the relative judgement tendency by removing its structural basis. Instead of presenting the witness with an array of parade members (i.e. exposing the witness to the suspect and a number of foils simultaneously), the parade members are ushered into a room individually and sequentially, and the witness is required to identify the perpetrator from the members of this sequence.

The results of five studies, comprising a total of about 15 experimental comparisons, show that the sequential parade is highly effective in combating the high false alarm rates found in so-called 'simultaneous parades'. It should be noted that accurate identification rates are not substantially affected by this manipulation, while false alarm rates are greatly reduced. A typical false alarm rate from a sequential parade will be below 0.10, as opposed to typical rates of more than 0.25 in simultaneous parades.

It is instructive to look at the results of a number of these studies (summarized in Table 4.7, below), specifically as a way of comparing the traditional, simultaneous police parade with the sequential parade.

---

[213] Later arguments by Wells and colleagues, discussed earlier in the chapter, suggest that the critical similarity relation is between foil and description of the perpetrator, and not between foil and suspect. The notion of a relative judgement strategy is unaffected in other ways by this revision.

Study		*Identification*	*Sequential identification parade*		*Simultaneous identification parade*	
			PP	PA	PP	PA
Lindsay and Wells, 1985		Accurate	0.50	0.65	0.58	0.42
		Mistaken	0.02	0.35	0.12	0.58
Lindsay, Lea & Fulford, 1991		Accurate	0.47	0.77	0.57	0.43
		Mistaken	0.07	0.17	0.20	0.37
Lindsay, Lea, Nosworthy et al. 1991	a	Accurate		0.93		0.70
		Mistaken		0.03		0.20
	b	Accurate		0.80		0.53
		Mistaken		0.03		0.20
	c	Accurate		0.67		0.23
		Mistaken		0.07		0.40
	d	Accurate		0.87		0.60
		Mistaken		0.03		0.13
Cutler & Penrod, 1988	a	Accurate	0.8		0.76	
		Mistaken	0.19		0.39	
	b	Accurate	0.41		0.47	
		Mistaken	0.21		0.43	
Sporer, 1993		Accurate	0.39	0.61	0.44	0.28
		Mistaken	0.61	0.39	0.56	0.72

PP = perpetrator present; PA = perpetrator absent. Lindsay, Lea, Nosworthy et al. do not report data for accurate identifications in PP lineups,[214] and Cutler & Penrod do not report results for PP and PA lineups separately.

Table 4.7 Mistaken and accurate identification rates in simultaneous and sequential identification parades.

The table shows two things very clearly. In the first place, the rate of accurate identifications is similar in both sequential and simultaneous parades[215] when the perpetrator is present in the lineup, but not when the perpetrator is absent from the lineup, in which case sequential parades have a considerable advantage. Thus, sequential parades are (almost) as likely to secure an identification when the perpetrator is present, and more likely to result in a correct rejection of the parade when the perpetrator is absent. More importantly, the rate of mistaken identifications differs dramatically across the types of parade structure: sequential parades lead to a greatly reduced rate of false identification when the perpetrator is absent. These results are quite exciting: they clearly indicate the superiority of the sequential parade. In addition, they are robust, and have been replicated in several studies. One of the most interesting replications is the study by Lindsay, Lea, Nosworthy et al., reported in Table 4.7: these researchers showed that sequential parades are much less dramatically affected by sources of bias, such as suspect-foil dissimilarity and suggestive instructions, than simultaneous parades. All the evidence points to the superiority of sequential parades; it is perhaps germane that police forces in South Africa adopt them on an experimental basis. They will involve little additional work in relation to that already required by the present parade structure and procedure.

[214] They found that the rate of accurate identification did not differ across simultaneous and sequential parades.

[215] Most of the research has shown that this conclusion can be accepted in individual studies with a great deal of statistical confidence. However, cumulated across studies it seems that the rate of accurate identifications is slightly higher in simultaneous parades. This is not surpising, and does not indicate an inherently greater ability of simultaneous parades to procure accurate identifications: guessing rates are higher in simultaneous parades, and the small difference in accuracy rates is probably due to inflation by guessing.

In this chapter, I attempted an overview of published psychological research on identification parades. This was perhaps a rather ambitious attempt, consequent on the perceived task of evaluating the contribution of psychology to a substantive legal problem. I will accordingly settle in this conclusion for observations relevant to one or two themes to be developed in later chapters of the thesis.

I have noted many interesting lines of psychological inquiry into the identification parade and its problems. The earliest of the lines of enquiry concerns the measurement of parade fairness with the mock witness task. The measures developed in this respect are promising, and show the potential of the empirical method to illuminate areas that have proved relatively inaccessible to the law. However, I have argued that the measures lack a firm basis in statistical theory, despite their elaboration in terms of uncertainty. In a later chapter, I will attempt to develop suitable statistical methods for each of the several measures considered in this chapter.

Psychological research on identification parades depends in large part for its justification on the prospect of application: this was the theme of Chapter 2. In this respect, it must be noted there is little evidence of direct interaction between research and law enforcement agencies at the level of research planning and execution. Such interaction, I argued earlier, is often one of the signs of a successful applied discipline. There are certainly some clear instances of contact and interaction between research and police practice - the adoption of the sequential identification parade in many U.S.A. precincts, for example - but there is little to suggest that the fate of identification parade research will be any different to other forms of applied psychology.

The most important observation that I wish to take forward from this chapter concerns physical similarity. In the first place, the notion that foils and suspects resemble each other is central to the legal conceptualization of parades. In the second place, it has rarely been addressed directly in psychological research. When it has, the measures employed to derive estimates of similarity are quite crude, but the results nevertheless suggest that similarity is crucial to the outcome of identification parades. Since few psychological studies have taken similarity into account, it may in addition be the case that much of the research is confounded by its unconstrained variation. One aim of the empirical work reported in later chapters is the development of a direct measure of facial similarity. A second, minor aim is to examine the robustness of some of the central parade research findings with respect to facial similarity.

Chapter 5 Face Representations and Theories of Face Recognition

In this chapter my aim is to raid the face perception and recognition literature for a suitable conceptualization of facial similarity. This follows naturally from the conclusions drawn in the two preceding chapters.[216] I will consider several possibilities, and argue for an adaptation of a particular approach to facial representation: this approach represents faces as eigenspaces of normalized intensity maps in the picture plane, and it allows - I suggest - a method of deriving a similarity space for populations of facial images. The route to this position is not uncomplicated, or brief; it takes us through much recent theoretical and empirical work, and there are several profitable stops to be made.

Theoretical aspects of face recognition

Until the early 1980's there was little formal theoretical work in the area of face recognition (Bruce, 1994). In the interim period, there has been a considerable amount of theory development and testing, and I will examine several of the more complete theoretical accounts in the next section of the chapter. There are a number of theoretical and meta-theoretical issues, though, that are relevant across specific models of face recognition, and which are of some importance to the empirical goals of this thesis. I will address these here, showing - I hope - that perceived facial similarity is best understood in a web of relations to more general problems in the face perception and recognition literature.

Configural vs. Featural processing

A long-standing issue in much face perception research is whether cognitive processing of faces is *configural* or *featural* in nature. Are faces perceived as combinations of separable features?, or are faces perceived as configurations, in which features are inherently

[216] In these chapters, I showed that legal requirements for conducting a parade assume that the similarity of parade members can be controlled, and I reviewed empirical evidence that suggests that variations in similarity can strongly affect parade outcomes. Although 'physical similarity' certainly encompasses more than just facial similarity, I consider only similarity of faces. I will provide some justification for this exclusion in the present chapter.

inextricable? This problem is not unique to the face perception area, indeed it is one of the fundamental perceptual questions. The notion that faces are processed as configurations is reminiscent of certain Gestalt propositions.

Although the literature does not support a unanimous conclusion, the weight of the evidence appears to favour the configurational view, or a medial position. Rhodes, Brennan & Carey (1987) note that the featural and configural positions are probably confounded, since a featural change automatically implies a configurational change. Thus, replacing Margaret Thatcher's nose with Edwina Currie's (on an identikit portrait) will not only yield a different single feature, but also a different configuration.

Feature saliency

The argument for the featural view of face processing derives from studies which address the saliency of different facial features in perception and recognition. There are several ingenious methods of exploring saliency, and much of the work is summarised in Shepherd, Davies & Ellis (1981). Nigel Haig, for example, has used image processing hardware and software to selectively degrade, embellish, and construct faces in terms of 'features'[217] (Haig, 1984, 1986a, 1986b).

The results of these studies appear to show that the upper face is more important[218] than the lower face, that internal features (eyes, mouth) are more important than external features (hair, head shape) for familiar faces, but less important for unfamiliar faces, and that hair length, hair coloration, and face thinness-fatness are also important cues (Shepherd et al., 1981). In general, the saliency studies that manipulate features show a fair amount of agreement in results, but studies which use subjective reports and verbal descriptions, and/or scaling methods, vary a great deal. The scaling studies, in particular, appear to produce results strongly dependent on the specific population of faces used as stimuli (Shepherd et al.).

The saliency studies have usually assumed that faces are processed featurally, but have rarely made a formal case for the featural model. As early as 1973, an intriguing result reported by Harmon suggested that a simple featural model must be false. Harmon showed that judgements of facial identity could be accurately made from representations consisting of as few as 16 pixels.[219]

---

[217] I emphasise the word 'features', since most featural studies have adopted a common sense notion that facial features are self evident - eyes, noses, chins, etc. This begs the question of what the important facial features are: large parts of the face are excluded as candidates, e.g. foreheads, cheeks. See Rhodes (1988).

[218] In the sense that recognition accuracy and latency are improved - different studies use different measures, so this is a very broad generalization. See Shepherd et al. (1981) for a discussion.

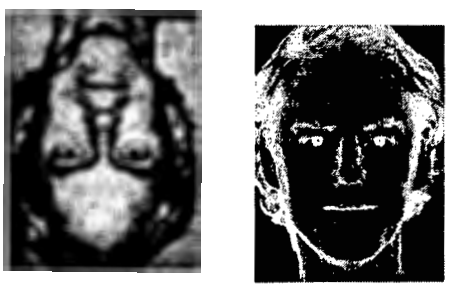
[219] *Pixel* = 'picture element'. Harmon showed subjects digital representations of familiar faces that were displayed as *k* blocks (or pixels) of differing light intensities. Thus, a face might be represented by 30 000 pixels of varying light

Bachmann (1991) found essentially similar results, using a different method. Since a face can be recognised from as few as 16 pixels, individual features cannot be of much importance: single features in a 16 pixel face will be represented by as few as one or two pixels, and the degraded nature of such a representation suggests the relative unimportance of single features. A similar argument can be made from the findings of studies in which representations of faces are spatially filtered to remove high frequency information.[220]

In a rigorous examination of the issue, Sergent (1984) carefully manipulated identikit portraits in terms of three features, and showed that latency on a same-different task was differentially affected by combinations of manipulations over simple manipulations (i.e. changing two features produced greater response latency than expected from the manipulation of each feature separately). In addition, a multi-dimensional scaling solution of similarity judgements yielded dimensions that were not independent, which suggests that facial features should not be considered in separation from each other (but should rather be considered in configuration).

Inversion effects

One of the better known early findings in studies of face recognition is the great difficulty that vertically inverted faces create for accurate identification and recognition. Yin's (1969) widely reported study first showed this, and the finding has been replicated many times (see Valentine, 1991a, for a review). Figure 5.1 demonstrates the effect, alongside another well known form of inversion (photographic negative inversion), which is also known to strongly affect recognition performance (Phillips, 1972).



Vertical inversion Photographic negative inversion

Figure 5.1 Two inversions known to affect the accuracy and latency of face recognition.

intensity (a high resolution display), or 100 such pixels (a low resolution display), or even 1 pixel. The question Harmon was interested in was how the progressive reduction of picture quality would affect judgements of face identity.

[220] The methods are almost identical in their consequences: pixellation in digital images achieves much the same result that spatial filtering achieves in analog images (Harmon, 1973).

Yin interpreted his results to imply a special route for face processing, but other authors have provided persuasive arguments against this view. Thus, Diamond and Carey (1986) showed that vertical inversion of exemplars of other categories of percept can equally disrupt recognition and identification, provided that the perceivers have expert knowledge of the exemplars.[221] The results from the inversion studies are often taken as evidence for the configurational view of face perception, since inverted and upright faces have identical facial features, and differ only in configuration. Valentine & Bruce (1988) dispute this interpretation, and provide data which is consistent with the idea that faces are 'mentally rotated' to an upright position from an inverted state: Increases in response latency are directly proportional to the required rotation, and recognition difficulties can be explained in terms of a multidimensional model of face encoding (see Valentine, 1991a).

Split half composite faces

Perhaps the most convincing evidence in favour of the configurational view comes from studies that use 'split' half faces. Young, Hellawell & Hay (1987) constructed face stimuli from photographs by combining bottom and top halves of photographs of different faces as i) composites, or ii) non-composites. An example of each, along with the original images, is shown below as Figure 5.2

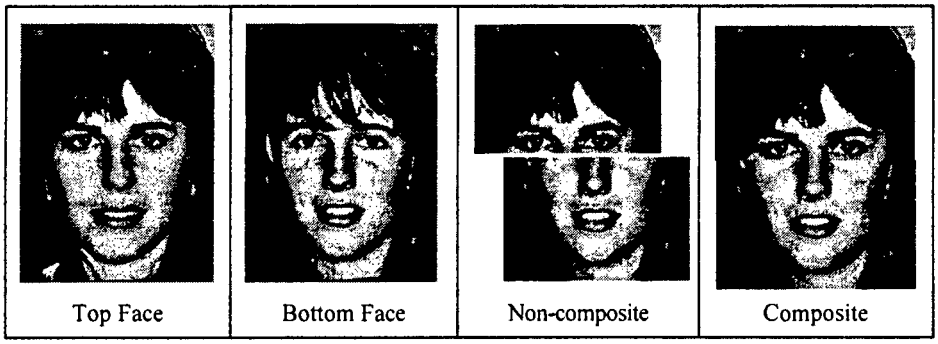


Figure 5.2 A split-half composite face, and its components (modelled on stimuli used by Young et al., 1987).

Subjects in Young et al.'s study found it comparatively easy to recognise the components of non-composite faces (i.e. who the bottom and top halves really belonged to), but made many errors when judging composites. This difference was not present when subjects were required to judge inverted faces. These findings were replicated with different composite constructions (e.g. internal features

[221] In particular, Diamond & Carey found that dog breeders made the same retarded recognition and latency responses to inverted representations of dogs that they showed to inverted human faces.

composited with external features), and with both well known and recently familiar faces. Facial components, or features, appear to take on different perceptual properties in different combinations.

These results make a strong case for the importance of configural cues in the recognition of faces. But, as Young et al. concede, highly distinctive features may frequently assist face recognition processes, and much more needs to be understood about the interaction of configural and featural cues.

First-order and second-order information

Diamond & Carey (1986), and Rhodes (1988), have suggested a formal way of conceptualizing the difference between featural and configural cues. They distinguish 'orders' of information in terms of category properties. First-order information refers to the properties that a single face has, taken as a perceptual object (e.g. eyes, ears, a nose, a 'face-like' configuration of features). Second-order information refers to the relational properties of facial components: all faces share a basic configuration, but there are many spatial differences between particular configurations, and these differences provide important information about the identity of faces. Rhodes (1988) suggests that it is possible to identify still higher orders of differentiating information, such as age and gender, but is not clear that extending the scheme offers any conceptual advantage.

The significant aspect of this 'ordering' of facial information for present purposes is the acknowledgement that configural information depends on the general nature of the category. Since the human face is an example of a natural category, the properties of the category derive from the population of category exemplars (see Rosch & Mervis, 1975).[222] This, I will later argue, is a crucial consideration in developing a measure of facial similarity.

Formal theories of face recognition

There are several formal theories of face recognition in the literature, and I would like to draw concusions from them for discussion later in the chapter. I will deal with two of the most widely researched and evolved theories: both of these are the work of Vicki Bruce and her colleagues in the United Kingdom.

Certain preliminary considerations are of some significance.

[222] Indeed, Wittgenstein (1950/1978) derived his notion of a 'family' resemblance from the configural variation of genetically related faces.

Familiar vs. unfamiliar faces

Most face recognition research prior to 1980 used unfamiliar faces as stimuli. These were usually presented to subjects briefly, and subjects were later tested on their ability to recognise the faces. Bruce (1979) noted the patent lack of ecological realism in this procedure, and suggested that it was analogous to Ebbinghaus' (1964/1885) use of nonsense syllables. In the interim period, much research has shown that familiar and unfamiliar faces deserve very different theoretical treatment. Thus, Ellis, Shepherd & Davies (1979) demonstrated that familiar faces are usually identified more easily and accurately from internal facial features, but that no such advantage holds for unfamiliar faces. Young, Hay, McWeeney, Flude & Ellis (1985) replicated this finding, and showed that it holds across variations in pose and expression.

Bruce (1994) reports work documenting the different effects of transformations on familiar and unfamiliar faces. Recognition accuracy for unfamiliar faces is more prone to changes in pose (e.g. from a 3/4 view to a full face view), than is the case for familiar faces. Indeed, familiar faces are fairly robust to such transformations. A similar result, obtained with very different methods, is reported by Read, Vokey & Hammersley (1990).

These studies suggest that familiar and unfamiliar faces receive qualitatively different processing, and furthermore, that familiar faces are more likely to be represented at the level of structural code than pictorial code.[223] This is an important observation for the empirical work reported in this dissertation, as eyewitnesses most usually identify people who are only recently, and slightly, familiar to them. Formal theories of face recognition are largely concerned with the processes underlying the recognition and identification of familiar faces, so I will not examine the theories very closely, but will take what I think useful forward to later chapters.

Information processing models

Most recent theories of face recognition have explicitly adopted an information processing approach. Recognition and identification is postulated to occur in a series of discrete steps, and is thus both stage- and time- dependent; at each step, incoming information is processed by functionally independent modules, and some (usually under-specified) central executive system co-ordinates the flow of information. Face recognition theories are often consciously modelled on exemplars of the information processing approach: Bruce & Young (1986) explicitly acknowledged that their model was based on the well known 'logogen' model made famous by Morton (1969).

---

[223] This distinction is clearer when considered in the light of the information processing model of Bruce & Young (1986).

The Bruce & Young model

The Bruce & Young (1986) model of face recognition - which developed out of earlier work by Hay & Young (1982) - is probably the most comprehensive single model of its type, and is also widely accepted as providing the most satisfactory account of extant research findings. Although the model concerns itself with recognition of faces, Bruce & Young argue that it could be considered more generally as a model of person recognition, since person recognition is largely dependent on face recognition. They provide evidence from various sources in favour of this proposition, and claim that the evidence from cases of neurological impairment is particularly compelling.

The model postulates that seven different types of information can be obtained in the act of perceiving familiar faces. In the language of the model, these types of information constitute *codes.*

The pictorial code

Faces are frequently encountered and learnt through a particular medium of representation. Many highly familiar faces, especially those of public figures and celebrities, are never encountered 'in the flesh', but are frequently seen on television, in magazines, and in newspapers. There is some evidence to suggest that such a pictorial code (i.e. a representation that is picture dependent) is implicated in face processing and recognition. Studies which expose subjects to particular representations of faces (e.g. photographs) report greater face recognition accuracy when subjects are tested with exactly the same representation than when they are tested with different representations of the same face (Shepherd, Ellis & Davies, 1982; Read, Vokey & Hammersley, 1990).

The structural code

Evidence for the involvement of a 'structural code' in face recognition is also strong. This is the sense in which people can see a photograph of a new face, and recognise a different photograph of the person as depicting that person: presumably, a code has been created which is invariant under certain transformations. There appears to be a difference in the structural codes for familiar and unfamiliar[224] faces: Familiar faces are encoded with respect to less changeable aspects of face. The nature of the structural code is unspecified in the model, but Bruce & Young take bearings on Marr's (1982) work on vision. Accordingly, they suggest that there are probably interdependent sets of descriptions at the level of structural code, allowing identification from single features (e.g. eyes) or from whole faces. Bruce & Young argue further that structural code is probably expression independent, and that the

[224] 'Unfamiliar' here denotes faces that have been seen before, but which are not well known. This is an unfortunate choice of word, but corresponds to usage of the term in the literature.

descriptions in the codes probably allow for different head angles. In short, structural code is *canonical.*

Visually derived semantic code

When we recognise faces, we are immediately able to make judgements about several personal attributes (for example, sex, age, occupation). We can think of people who resemble the person in question, we can make judgements of relative beauty - in short, we tend to derive a great deal of semantic information about the person from visual recognition. In the model, this is referred to as visually derived semantic code.

Identity specific semantic code

Recognition of particular faces also provides immediate information about the recognised person. Typically, when the person's face is highly familiar, we gain access to information about where she lives, who she is married to, what she does for a living, and so on. This information bears very little relationship to the physical form of the face, and is referred to in the model as identity specific semantic code.

Name code

Although one might be tempted to subsume information about a person's name under the previous code, Bruce & Young make a strong case for the existence of a separate name code. In particular, they propose that it is an output code, which allows a name to be generated, and passed on to 'output devices' (like a language production system). Some of the most compelling support for such a code includes i) the well known everyday difficulties we (humans) have in producing a name, although we can simultaneously remember other attributes of the person in question; and ii) evidence from neurological disorders, to wit apraxias, where subjects are completely unable to generate names, but other functions are intact.

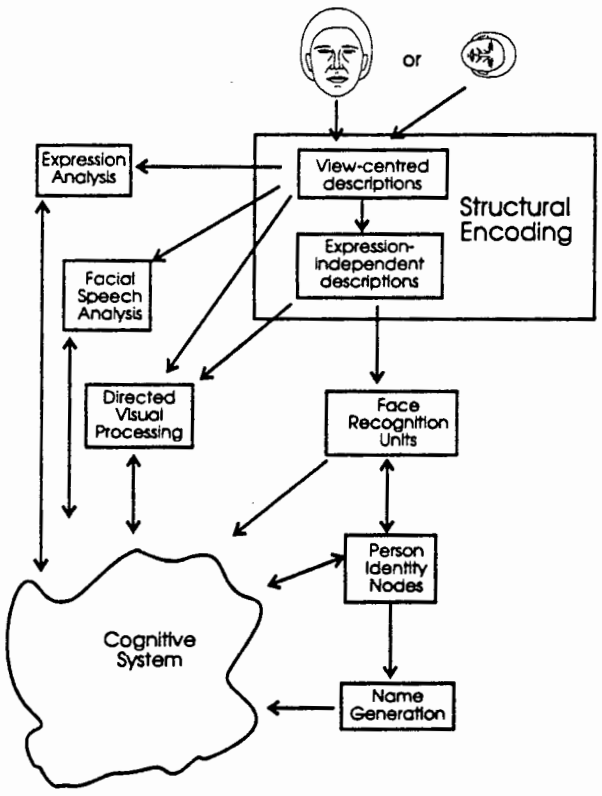
Other codes postulated by the model, but of less significance, are the *facial expression* code, and the *facial speech* code.

A diagrammatic representation of the model is presented below, and the role of the seven codes in the functioning of the model is evident from the diagram.

Two additional postulates of the model are of considerable significance. These are i) the existence of logogen-like devices, known as 'face recognition units' (FRUs), and ii) the existence of 'person identity nodes'.

Face recognition units (FRUs) and person identity nodes (PINs)

Face recognition units receive input from encoders, which transform code obtained from the visual system into a suitable form. Each FRU contains structural code describing one of the faces known to the individual (there are therefore at least as many FRUs in the system as there are known faces), and is activated by the input it receives. The strength of this activation, and the strength of the signal that the FRU sends to the cognitive system, will depend on the similarity of the input stimulus code to the code stored in the FRU. Bruce & Young posit that FRUs don't operate in a threshold firing mode, since even strong degrees of resemblance will not necessarily lead to 'recognition', particularly when the person is a 'look alike' and context evaluation indicates a mismatch. The FRU can be 'primed' by PINs, e.g. in anticipation of someone we know we are about to meet. (PINs contain identity specific semantic information). The PINs are the site at which person recognition occurs; and as such can be activated by FRUs, or by voice codes, or even by other codes. FRUs, on the other hand, must be fed visual information.



(Reproduced, with minor modifications, from Bruce & Young, 1986, p 312).

Figure 5.3 The Bruce & Young (1986) model of face recognition.

The model is sequential in nature: name codes must be generated after identity specific information, for example, and face recognition units can only fire after visual input is structurally encoded. This does not exclude the possibility that there is an interactive flow of information (indeed, double headed arrows in the diagram indicate where this is explicitly postulated), nor that later stages may influence processing at earlier stages. There is interlinkage between the 'codes' and the processing stages. Thus, name codes appear to be accessed through identity specific semantic codes at the level of the PINs. The model is supported by a host of findings, some deriving from empirical studies specifically designed to test it (e.g. Ellis, 1992; Roberts & Bruce, 1988; Valentine & Bruce, 1986b), some from diary studies of everyday facial memory problems (Young & Hay, 1985), and a good deal also from neurological knowledge of difficulties encountered in recognition and identification of faces (Bruce & Young, 1986).

It is of little benefit to the present chapter to attempt an evaluation of the model. Comprehensive analyses can be found in Ellis (1992), Hay & Young (1991), and Bruce (1988), and the addition of a 'back-end' name generator is proposed by Valentine, Bredart, Lawson & Ward (1991). In the most recent developments, the model has suffered the fate of many information processing models of cognition, namely its virtual replacement by a PDP (parallel distributed processing) equivalent. I will look briefly at the replacement in the following section of the chapter, but wish to identify two aspects of the information processing model that will be important when I consider possible measures of facial similarity.

The face recognition model suggested by Bruce & Young has a visual 'front-end' which is left largely unspecified. This is not an omission, indeed Bruce has at several places acknowledged the importance of this pre-processing stage (Bruce, 1992, 1994). What interests me is the postulate that the first stage of visual processing after the front-end involves the extraction of a 'structural', or 'canonical' code. The model has nothing specific to say about such a code, except that it is likely to be invariant to transformations such as pose and expression, and that each face is elaborated in terms of structural code. I suggest that an analogue of such a canonical code is a method of representing sets of digitized facial images in an eigenspace, recently developed by Kohonen (1984), Sirovich & Kirby (1987), and O'Toole and colleagues (O'Toole, Milward, & Anderson, 1988; O'Toole & Thompson, 1993; O'Toole, Abdi, Deffenbacher & Valentin, 1993). This goes somewhat further than suggested in the Bruce & Young model, which proposes only that *individual* faces are represented in terms of a structural code. I will later make the argument that the representation of populations of faces in terms of common reference axes (or eigenvectors) provides us with a basis for measuring facial similarity.

The second aspect of the model that I think is directly relevant to present concerns is the differentiation between more- and less- familiar faces. Face recognition units exist for each known face, but structural code in the units is 'better developed' for more familiar faces, and is elaborated in terms of stable, less changeable facial properties. I think that the representation of faces by a basis set of eigenvectors might give clearer meaning to these propositions.

Instance based models

Just as the information processing models of the 1960's and 70's swept away previous theories of cognition and learning, so parallel distributed processing models are sweeping away those rooted in the information processing paradigm. These models, known variously as 'PDP', 'instance based', 'exemplar based', 'associationist', or 'neural networks', prosper because of their remarkable simplicity, and their equally remarkable success at simulating complex behaviour (see McLelland & Rumelhart, 1985, 1986, for examples). In particular, they appear to explain many examples of cognitive behaviour which require expert knowledge of naturally occurring categories without recourse to the complex systems of rules that information processing approaches postulate. The proferred explanation posits only experience of a discrete number of instances of the category, and a general associationist model of learning.

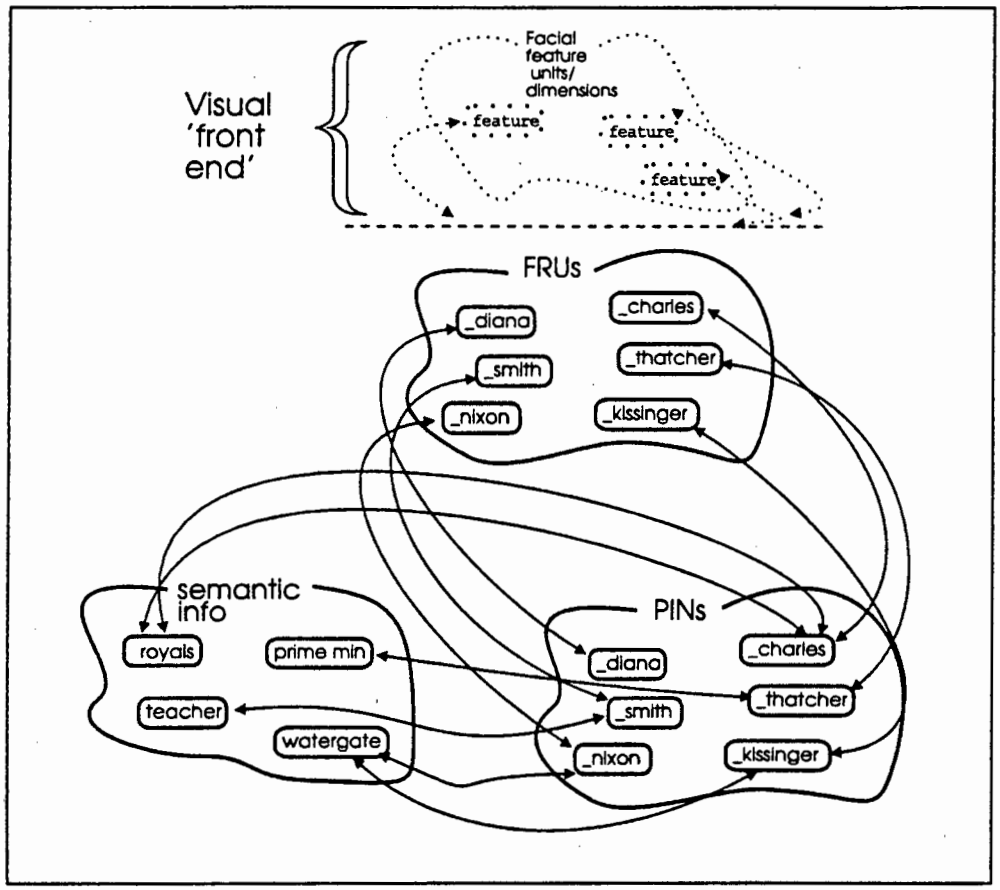
In the case of face recognition, there are several PDP models in the literature. I will focus only on the Burton & Bruce (1990) model here,[225] but will refer to the auto-associative model developed by Kohonen (1984) a little later in the chapter.

The Burton & Bruce model

The 'Burton & Bruce model', developed by Burton & Bruce (1990), and extended by Bruce, Burton & Craw (1992), Burton (1992), and Bruce, Burton & Walker (1994), is based closely on the Interactive Activation and Competition network (IAC) model first researched and published by McClelland & Rumelhart (1981). The model, represented below as Figure 5.4, has hypothetical units or 'neurons', which are connected multiply to other such units. These connections can be excitatory or inhibitory, and are bi-directional. Units are grouped in 'pools': There are three of these in the model, namely a pool of FRUs (face recognition units), a pool of PINs (personal identity units, or nodes), and a pool of semantic information units.

[225] Other models include those suggested by Schreiber & Rousset (1991) and by Li & Psaltis (1993).

Information enters the system via a visual or perceptual 'front-end': a rudimentary version derived from Burton & Bruce (1990) is depicted in the diagram, consisting of 'feature units' connected in pools. Activation spreads from the front-end to the FRUs, and this is passed on to units in other pools according to the level of activation of each unit, and the weighting of the connection between the units. Familiarity decisions are taken at the level of the PINs, partly as a function of the activity passing there from the FRUs. i.e. this activity is 'combined' with information from other routes, to achieve identification of the person rather than the face in isolation. PINs are multi-modal nodes - they are the entry point to semantic system - but FRUs act as unimodal nodes, pooling information over a set of more primitive feature analyzers. There is excitation between pools, and inhibition within pools.



(from Burton & Bruce, 1990; modified to include a 'visual front-end').

Figure 5.4 Burton & Bruce's IAC face recognition model

Many well established findings in the face recognition literature can be explained in terms of the functioning of this model, including priming effects, covert recognition, and naming difficulties (Bruce, Burton & Craw, 1992). Some of these findings were very difficult to explain in terms of

information processing models like that developed by Bruce & Young, discussed earlier in the chapter.

The model has two significant failings. In the first place, there is no learning mechanism - it is a steady-state model, but one of the most important aspects of human face recognition is the learning of new faces. The auto-associative models posited by Kohonen (1984) and O'Toole & Thompson (1993) though, provide a learning mechanism in the form of the so-called delta or Widrow-Hoff rule. This will be an important consideration when I evaluate representational schemas that might sustain a similarity metric.

The second significant failing is the rudimentary elaboration of the perceptual interface. This problem is shared by information processing models, and its satisfactory resolution remains a major challenge for face recognition research.

Facial representations

In both face recognition models considered above, the lack of an adequate perceptual interface or 'visual front-end' is a significant impediment. Similarly - or perhaps in consequence - neither of the models specifies a representational schema. There are several interesting ideas for such a schema in the literature, though, and I will review these in this section. The multidimensional schema suggested by Valentine (1991a) is, in particular, an important foundation of the similarity metric to be advanced later.

Surfaces in 3D

Much face research has proceeded as if two dimensional pictorial representations of faces are adequate physical specification. In many of the studies that test theoretical claims, subjects are given photographs of faces to learn, and sometime later are tested for recognition ability. But faces grow on heads, and these are clearly three dimensional objects. Several researchers have demonstrated that information about the three dimensional nature of faces is important to face recognition ability.

Pearson and colleagues (Pearson & Robinson, 1985; Pearson, Hanna & Martinez, 1990; Pearson, 1992) showed that the two-dimensional edge detectors favoured in Marr's (1982) theory of vision fail to provide easily recognisable representations of faces. They have developed two- and three-composite operators which capitalize on a three dimensional model of 'seeing' a face from a position nearly co-incidental with the surface normal to the face. These operators produce automated sketches of faces for use in low-bit-rate video communication for deaf people, and are strikingly similar to sketches produced by trained artists (Pearson, 1992). Bruce (1994) argues that this research

demonstrates the importance of three dimensional representations to an adequate understanding of face recognition processes.

Bruce & Healey (1991) attempted to explore the role of three dimensional information directly, by using a laser scanning device to create facial images, thus rendering them devoid of surface pigmentation. Subjects were asked at the test phase of the study to identify faces that they had been exposed to at the beginning of the experiment. Subjects found this a very difficult task, though, and showed much greater recognition accuracy for photographs of the same faces.

In later research, Bruce, Hanna, Dench, Healey & Burton (1992) pursued this issue, using the cartoon generator developed by Brennan (1985). Previous research had shown that veridical line drawings of faces are more difficult to recognise than photographs. Bruce et al. tested the notion that 'mass' (low frequency information) is critical to the advantage of 3-D representations. The cartoon generator was used to create line drawings from photographs, and subjects were compared for their performance on recognition tests using either photographs or line drawings as stimuli. Recognition performance on the cartoons was good - indeed, comparable to recognition performance on the photographs.

Further evidence suggesting the importance of three dimensional information comes from a study in which extensive sets of facial measurements were taken and used to build statistical models predictive of the sex of faces (Burton, Bruce & Dench, 1993). Here, three-dimensional measures were more highly predictive than two-dimensional measures, and appeared also to be more important.[226]

Most recently, Bruce, Coombes & Richards (1993) have attempted to specify a general representational schema for faces, which takes surfaces in three dimensional space as its primitive objects. In this schema, a face is represented as a composite of up to eight surface types, namely peaks, ridges, saddle ridges, pits, valleys, saddle valleys, flat regions, and minimal regions. Faces can be compared in this scheme according to the percentage of the facial surface each surface type occupies, and according to the combination of surface types required to adequately represent the face. Preliminary empirical research using this schema shows that it is capable of describing facial distinctiveness and typicality. It is also a satisfying schema on a priori grounds, since its representations are invariant to viewing perspective or pose. However, the authors concede that it may not be a good model for the way humans perceive and recognise faces. This is not a critical issue for the present research, since the goal is to develop a measure of facial similarity: This measure should correspond to perceived similarity, but functional equivalence is not necessary. From a practical point of view, the equipment required to produce a facial representation in terms of 3D

[226] The evidence in favour of this latter proposition, though, is that three dimensional measures entered the model equation first (a stepwise procedure was used). This is not a particularly good way of judging the relative importance of variables (Howell, 1992).

surface primitives is expensive, and bulky.[227] In addition, the schema does not explicitly account for variation within populations of faces, and this is a significant omission: in the next section, we will see that facial distinctiveness and typicality are important variables.

The representational schema I argue for later in the chapter does not attempt to produce three dimensional representations. This is clearly problematic, given the discussion immediately above. However, the schema takes three dimensional information into account implicitly, and I will suggest that this is adequate from a pragmatic point of view.

Norm based coding models

Research on inversion effects in face recognition clearly points to the importance of knowledge of intra-category variation (see page 102 of this chapter). There are several representational schemas which attempt to incorporate intra-category and intra-population variation, and these are perhaps the most promising approaches for deriving a similarity metric. In order to discuss them adequately, I first briefly review research on the role that facial distinctiveness and typicality play in face recognition.

Typicality and distinctiveness of faces

The central findings are twofold. Typical faces are easier to classify as faces, but more difficult to identify individually; they also attract a greater number of false alarms in recognition tasks. Distinctive faces, on the other hand, are easier to identify than typical faces.[228]

Valentine & Bruce (1986a), for example, showed that subjects classified typical faces more quickly than distinctive faces in a 'jumbled face' task,[229] and identified distinctive faces more quickly than typical faces in a recognition task. Similar results were obtained earlier by Light, Kayra-Stuart, & Hollander (1979).

Facial distinctiveness and typicality scores are usually obtained from ratings of subject-judges, and are therefore strongly dependent on the set of faces used,[230] and on the set of faces with which the judges are experientially familiar. Shepherd, Gibling & Ellis (1991) attempted to counteract the first

---

[227] The equipment consists of a laser scanner, several precisely aligned optical mirrors, and a very powerful computer.

[228] I mean 'easier' here in the sense that they are identified more quickly, and in the sense that they are identified with greater accuracy.

[229] Faces are presented intact or as jumbled composites of face parts. Subjects are required to indicate as quickly as possible whether the face is intact or jumbled.

[230] This dependency is greatest when a ranking task is used, i.e. the set is ranked on typicality and distinctiveness.

of these problems by using a set of 240 faces, selected from a larger database of 1000. They demonstrated a strong effect of distinctiveness on recognition performance, and in addition showed that the effect held over the maximum investigated period of delay, namely one month.

'Distinctiveness' and 'typicality' are conceptualized in the literature according to their ordinary linguistic usage, but a study by Vokey & Read (1992) suggests a more complicated state of affairs. These authors conducted a study in which subjects rated typicality, familiarity, attractiveness, likeability and memorability of a number of faces. The ratings were factor analysed, and two components identified - an overall memorability factor or component, and one on which the other variables loaded, named the 'general familiarity' factor. Thus, typicality can be decomposed into two factors: context-free familiarity, and memorability. These factors were found to be additively predictive of face recognition performance; memorability enhanced discrimination, whereas familiarity impeded it. The results of this study suggest the need for theoretical elaboration of the notions of distinctiveness and typicality.

Several studies have made attempts at such elaboration in terms of facial prototype models. (A prototype in this sense is the 'best' exemplar of a category). Typicality and distinctiveness can be defined in terms of a notional 'distance' from the prototype, or in terms of the frequency with which the type occurs.

Facial prototype formation

Valentine & Bruce (1986c) outline a 'facial prototype hypothesis' in which prototype formation is some sort of averaging process, based on all faces experienced by the subject. The averaging process could be based on feature frequency information, and the prototype would then be the face composed of most frequently occurring features. Alternatively, there could be several prototypes, based on a similar sort of averaging process.

Malpass and Hughes (1986), however, point out that there are several other potential cognitive mechanisms that might function to form facial prototypes. They identify three rival models.

1 an attribute frequency model. Prototype formation occurs by taking the modal feature value for each of a set of faces.

2 a value averaging model. Prototypes are formed by extracting the mean value for each feature dimension.

3 an interval storage model. Types 'activate' not only the particular feature values corresponding to their features, but also the feature values directly adjacent to these. The prototype is thus stored as a set of features with associated average intervals.

Malpass & Hughes tested these models against each other by varying Identikit faces, and the results appeared to support the attribute frequency model. However, the averaging models were investigated by treating the dimensions along which identikits were manipulated as 'scales', whereas more direct measurement of faces might constitute a better test.

Whatever the dispute about the appropriate prototype model, several studies using very different methods have shown that there are strong prototype effects in face recognition. Solso & McCarthy (1981), for example, conducted an experiment in which subjects were shown identikit faces which varied in terms of number of shared features. These features were taken from a prototype face, which was not shown in the presentation stage of the experiment. Faces shared 1, 2, or 3 features. Subjects then completed a recognition test composed of 'old' and 'new' faces,[231] and indicated a degree of confidence for each choice. Subjects responded with greater confidence to the (previously unpresented) prototype face than to the other faces. This suggests that subjects formed a prototype from the presentation set.

Laughery, Jensen & Wogalter (1988) constructed sets of photo-lineups from Identikit composites and Macamug images,[232] and tested the ability of subjects to identify a face previously presented to them. Sets of faces were constructed from identikit prototypes, in a manner similar to that used by Solso & McCarthy (1981). Lineups were then constituted so that distractors differed in one feature from the prototype, and also differed in one feature from each other. Both prototypes and non-prototypes were used as targets and as distractors. Prototypes were recognised more frequently than non-prototypes. The authors suggest that this finding spells a warning to lineup constructors: attempting to maximise distractor - suspect similarity may have the unintended consequence of making the suspect a prototype, and therefore suggesting his identity to witnesses.

Bruce, Doyle, Dench & Burton (1991) investigated prototype formation in a series of carefully devised experiments, using prototypes and exemplars created with Macamug software. They report several important findings: i) even tiny changes in facial configurations[233] are important to recognition success; ii) prototypes seem to be learned by exposure to very similar faces; iii) prototype faces are preferred over 'extreme' exemplars, even when the prototype is not part of the trial set, and the extreme exemplars are. However, these results - and those reported in similar studies - do not necessarily imply a prototype learning model. Many PDP models are able to simulate category learning behaviour, and these models do not include any specific rule-learning mechanism: their

[231] That is, faces shown them during the presentation phase, and faces not shown them in that phase.

[232] 'Macamug' is a specialised face-image software package, and runs on Apple Macintosh computers.

[233] Moving the upper or lower part of the face up or down three pixels is an example of a 'tiny' change.

power is to show that mere exposure to a sufficient number of exemplars can produce the same behaviour as a rule-learning model.

These 'prototype studies' do not take us any further, in a practical sense, to the desired measure of facial similarity, but they do show that faces lend themselves to representation in a system with an inherent relational metric. More importantly, they show that human face perception and recognition is highly sensitive to such an underlying metric.

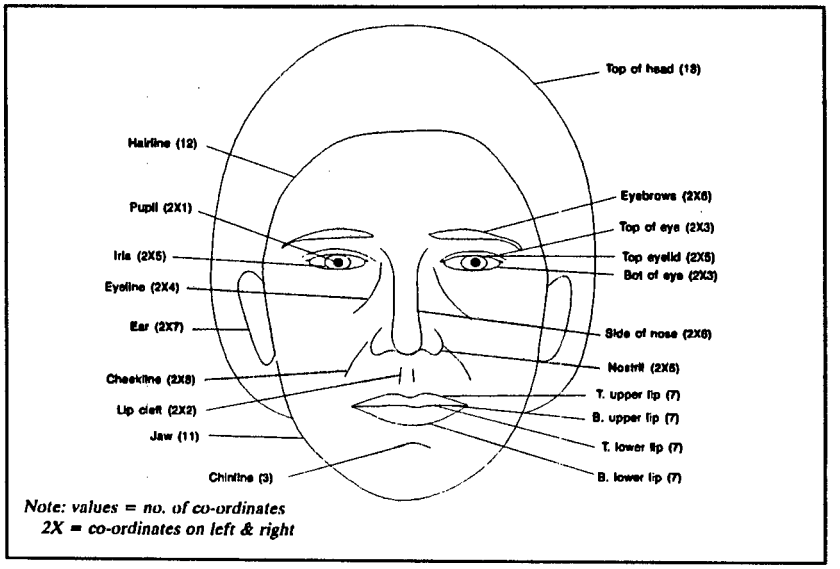
Caricature models

A different approach to the representational problem is taken by a set of studies which postulate that faces are encoded as 'caricatures'. Caricatures are defined as representations exaggerated metrically from an average, or norm, and the idea underlying this schema is that such exaggeration manages to capture the 'essence' of the face (the so-called 'superfidelity hypothesis'). Much of the evidence in favour of this representational schema derives from a demonstrated recognition advantage for caricatured faces over veridical portraits.

Although early studies did not support the 'superfidelity' hypothesis (Tversky & Baratz, 1985; Hagen & Perkins, 1983), Rhodes, Brennan & Carey (1987) argued that this was because comparisons were usually made between line-caricatures and photographs, and because face stimuli were not well controlled in several other ways (for example, the photograph and caricature might show the same person in different poses). They used a 'caricature generator' developed by Susan Brennan (Brennan, 1985) to create caricatures, veridical images, and anti-caricatures, thus exercising greater control over the stimuli. This caricature generator uses a representational scheme which is worth outlining at some length. Brennan argues that caricature proceeds by exaggerating the differences between two representations; usually between the object to be caricatured and some normative representation (e.g. an average, or normative face). Applying her theory of caricature to the example of faces, Brennan suggested that representations of faces could be produced by digitizing a number (say N) of discrete and salient points on any particular face from some two dimensional representation like a photograph or a computer-scanned bitmap. In this way, the Euclidean distance between two faces could be calculated by i) considering each of the faces as a vector in N-dimensional space, where each of the N co-ordinates of the point corresponds to a digitized point in two dimensional space; and by ii) determining the Euclidean distance between the two N-dimensional vectors, i.e. distance $= \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_n - B_n)^2}$, where A_i (or B_i) is the i^{th} dimension (x,y co-ordinate) of face A (or B).

In addition, it should be possible to estimate the distinctiveness of any particular face by finding the difference between the N-dimensional representation and a population-normative N-dimensional

representation (which itself could be estimated by averaging a large number of appropriately sampled representations). Brennan suggests that co-ordinates should be collected according to sets of features, and that co-ordinates chosen should correspond to points of inflection. Brennan's schema is reproduced below as Figure 5.5



Co-ordinates for the normative face reported in Dewdney (1986) are used to generate the face.

Figure 5.5 Brennan's scheme of fiducial points for representing faces in the picture plane.

Using the caricature generator to create stimuli of faces, Rhodes et al. (1987) showed that subjects were quicker to identify caricatures of familiar faces than they were to identify anti-caricatures[234] or veridical portraits of the same faces. There was no difference, however, in recognition accuracy across these conditions.

In further studies, Rhodes and her colleagues have extended the work on caricature effects. Thus, Rhodes & Moody (1990) used caricatures of unfamiliar faces as stimuli and tested recognition of these against anti-caricatures and veridical line drawings, along with different encoding instructions. They did not find an advantage for caricatures, but did find that subjects were able to reject caricatures of faces not previously seen more quickly than veridical drawings. Rhodes & MacLean (1990) replicated the caricature advantage (again, only for response latency and not recognition performance), using birds as stimuli, and ornithologists as subjects. Interestingly, the advantage was only obtained for expert ornithologists (not for less expert 'avian knowers'), and only more typical examples of bird species facilitated such an advantage.

[234] Anti-caricatures are defined as exaggerations *towards* the norm, whereas caricatures are exaggerations *away* from the norm.

Although the studies conducted by Rhodes and colleagues have repeatedly shown an advantage for caricatures over veridical portraits in terms of response latency, they have not shown a recognition advantage (i.e. caricatures of faces are not recognised with greater accuracy than veridical portraits). Several studies by other authors, though, have shown a recognition advantage for caricatures. Mauro & Kubovy (1992) used a caricaturing method in which distinctive features were exaggerated, and they showed that these were recognised more frequently than the veridical portraits on which the caricatures had been based.

Benson & Perrett (1991a) extended Brennan's caricaturing technique for use with high resolution bitmaps, and replicated both the caricature advantage for response latency, and the recognition advantage. The details of their technique are of some significance to a later discussion.

In their scheme, faces are represented in a two dimensional photographic plane. Key points are identified on a digitized photograph of a face, and are used to create a point-determined caricature of the face, in exactly the same manner as in Brennan's technique. Two sets of triangles (or triangular tesselations) are then constructed for each of the representations (veridical and caricature), by joining certain points with straight lines. (Figure 5.6 shows a simplified example of an image tessellation). The triangles map onto each other in the two representations, since they are formed by connecting key points, and the faces share these, albeit in a different spatial configuration. In the veridical face, each triangle contains a set of pixel values, since the face is a digitized image. These values are mapped into the corresponding triangle on the caricatured face, using bilinear interpolation and an algorithm reported in Benson and Perrett (1991b). The result is a caricature of near-photographic quality.

Figure 5.6 Sample triangular tessellation of a face image.

The caricature advantage appears to be manifested more clearly when these high resolution caricatures are used as stimuli. Benson and Perrett (1991a) showed that caricatures not only facilitated recognition latency over veridical stimuli, but also increased recognition accuracy.

Approaches to caricature generation rely on the representation of faces in a multidimensional space, although this reliance is perhaps not crucial to the techniques. They certainly show that such an approach has useful and powerful implications. Several authors have suggested models of face representation and recognition that capitalize on the notion of multidimensional space, and I will later argue that this notion may provide a useful foundation for a metric of facial similarity. In the next section of the chapter, I review some of the approaches that have pursued this line of enquiry.

Multi-dimensional space models

The idea common to face recognition models that use the notion of multidimensional space is that a population of faces can be represented in a space subtended by a number of 'primitive' dimensions. Each face is represented in this schema by a set of weighted basis vectors or dimensions. Such a conceptualization gives quantitative meaning to the notion of a 'normative' or 'average' face, and it also gives quantitative meaning to several important concepts, such as 'facial distinctiveness', and 'facial typicality'.

Although several authors have discussed the possibility of an explicit multidimensional model of face recognition (e.g. Rhodes, 1988[235]), Tim Valentine (1991a, 1991b; Valentine & Ferrara, 1991; Valentine & Endo, 1992) has provided the most comprehensive discussion, and a working example of such a model. I will accordingly restrict my discussion to a consideration of his contributions.

Valentine suggests that faces can be regarded as compositions of 'dimensions', which are attributes that exist for each face at some positive, mutually exclusive value. Each face can therefore be represented as an n-tuple, which is the set of values that the face has on each of the dimensions. The exact nature of the dimensions is not important - these may turn out to be physiognomic features, like noses, chins and ears, but they may well be something more abstract. Indeed, any discernible aspect of faces that serves to discriminate faces is a candidate. Faces are postulated to be normally distributed around the central tendency, or origin of the dimensions, and are therefore represented in the space with decreasing density as the distance from the origin increases.

[235] An implicit multidimensional model also underlies Brennan's 'caricature generator; outlined earlier in the chapter. The work on facial prototype effects also appears to assume such a model.

This model has implications for the recognition process. Each stimulus face will be encoded in an n-dimensional form, with some amount of noise. A decision process is therefore required to determine whether or not the stimulus matches the vector (or point) for a known face. This process is assumed to depend on i) the amount of noise in the encoding; ii) the distance between the location of the stimulus and its nearest neighbour; and iii) the distance between the location of the stimulus and the next nearest neighbour.

There are two possible geometric (or mathematical) interpretations of such a model. In the first, it is assumed that faces are represented as vectors in the space, extending from the origin outward. In the second, it is assumed that faces are represented as points in the space. The differences between the models are not trivial, and require some discussion.[236]

The vector coding model

In the vector coding model, the existence of a 'normative' representation is required. This means that faces need to be explicitly averaged over instances, and that such a normative representation be used in comparative judgements about typicality and distinctiveness. But there are identifiable sub-groups of faces, e.g. male and female, black and white, and it is not clear whether different normative representations will exist for different subgroups.

The vector, or norm coding model can provide an explanation for some of the empirically observed phenomena in the face recognition literature. That distinctive faces are recognised more quickly than typical faces is due to the difference in the ratio [error/distance to the origin] between distinctive and typical faces. Typical faces encoded with a certain amount of error will be confused with many other faces that are of a similar distance to the origin, whereas distinctive faces will not, since there are fewer faces of a similar distance to the origin. A similar explanation can be offered for inversion effects: inversion is bound to result in a particularly 'noisy' representation of the stimulus; this 'noisy' representation will be used to search the space, but will not be able to adequately discriminate faces that are similar to the stimulus, but not identical.[237] Recognition of faces will consequently be very poor.

---

[236] The differences derive directly from the two possible ways of describing an entity in n-dimensional space; i.e. as a vector v, extending from the origin to the point, or as the n-tuple representing the point, $(x_1, x_2, ...x_{n-1}, x_n)$. At the level of the recognition model, however, the crucial difference is that in the vector coding model an explicit norm is assumed, which is not the case for the instance based model.

[237] The mathematical definition of an 'open ball' could be used to give this idea clear quantitative meaning.

The instance based model

In the instance based model, the existence of a 'normative' representation is not required. A normative representation may exist, in the sense that all instances could be averaged to produce such a representation. However, a normative representation plays no role in the recognition process. The prototype effects discussed earlier in the chapter do not pose a problem for this model, since many distributed processing models show that is possible to produce such effects without assuming rule- or norm- determined processes (see Bechtel, 1991).

Typicality and distinctiveness effects are explained in this model by the relative density of points in regions of the space, and by the degree of error involved in encoding facial images. A distinctive face will be easier to recognize because a 'circle of confusion' at a region of low point density will include fewer points than one at a region of high density.

It is difficult to distinguish the models at the level of prediction. Valentine suggests that they offer indistinguishable explanations of typicality and distinctiveness phenomena, but that it may be possible to test them for their explanations of the 'other-race' effect.[238] The exemplar model posits that other race faces will be encoded in terms of the same dimensions as own-race faces, but less accurately. This means that they will be more densely distributed in the space (since the dimensions are not optimal for discrimination), and recognition will therefore be less sensitive. Importantly, it makes an additive prediction for the effects of distinctiveness and race in combination (one will have greater difficulty recognizing other race faces, but relative to this, distinctive faces in each race group will show the distinctiveness effect). The norm-based model makes slightly different predictions. If it assumes that both own and other race faces are encoded with respect to the same norm, in the same axis space, then although own race faces will be distributed at a full complement of different angles around the norm, other race faces will be distributed in one direction. It is not necessary for there to be a difference in exemplar density. Since other race faces share similar angle components, it is difficult to differentiate them in comparison to own race faces which will have wide differences in angle components. But this model cannot accommodate the additive effects of race and distinctiveness, unless it is assumed that other race faces are coded with respect to a separate norm. Valentine (1991a) suggests that this would be an unparsimonious assumption, but Benson & Perrett (1991a) argue - apropos of their caricature procedure - that the existence of different norms for different groups of faces is sensible if cognitive processes involved in face recognition resemble the

[238] The 'other-race' effect in the face recognition literature is the well-established finding that white subjects are better at recognizing faces of their own 'race' or group, than faces of other 'races' or groups, and vice versa. (Malpass & Kravitz, 1969; Brigham & Malpass, 1985). This difference is not vast, but is reliable, conferring an advantage of about 10% on the recognizing group.

caricaturing described by their model. In later research, Valentine & Endo (1992) concluded that empirical evidence from the face recognition literature supports the instance model more than it does the norm-coding model.

What is important for present purposes is the power of the multi-dimensional approach: it provides a representational model for face recognition - even if the nature of the dimensional basis is unspecified - and it also explains several of the well-established empirical findings. Significantly, it has a similarity metric embedded in it, where the similarity of two faces is a monotonic function of the distance between them (in terms of the n-dimensional representation, and a distance measure[239]). This is one part of the measure to be suggested in the present piece of work; the other concerns the dimensions that constitute the basis of the face space. In order to arrive at this point, it is useful to dwell on facial similarity measures previously used in the face recognition literature.

Previous attempts to measure similarity

It is useful to divide previous measures into categories; these are identified below.

'A priori' techniques

In these techniques, researchers use some criterion that is presumed to distinguish faces on the basis of their similarity. Thus, Patterson & Baddeley (1977) created groups which ostensibly differed in the facial similarity of their members, by using photographs of people from very different social categories: to wit, actors (low similarity, since there is no defining attribute of this group to ensure similarity), and soldiers (high similarity, due to common characteristics determined by shared age, haircuts, etc). Malpass & Devine (1983) created lineups of varying similarity by assigning individuals according to their height, weight, hair colour, hair length, and eye colour.[240] Laughery, Fessler, Lenorovitz & Yoblick (1974) operationalised similarity in terms of a set of trials in which faces were paired, where subjects were required to discriminate old from new faces. The proportion of mistaken ('old') responses was used to define similarity. In a second part of that study, similarity was operationalized in terms of physical characteristics (hair colour, age, etc.): these characteristics were used to construct a matrix, and similarity was defined as the number of shared characteristics.

[239] There are several possibilities: the Euclidean distance, the Minkowski power metric, and the city-block distance metric are usually considered in psychological work on similarity (see, for example, Shephard, 1958, 1987).

[240] Although it should be noted that Malpass & Devine also obtained independent ratings of the similarity of lineup members.

The weaknesses inherent in these types of technique include i) the untested nature of the assumptions used to determine similarity, and ii) their impreciseness. Although it might be reasonable to assume that groups of actors and soldiers will show different variability in facial similarity, this is a gross division, and of little use in most situations where similarity needs to be measured or manipulated.

Rating techniques

Most psychological studies that attempt to measure facial similarity do so by obtaining ratings of faces from independent subject judges. Bruce (1979) required subjects to rate stimulus faces in relation to target faces on a 4 point scale; in Milord's (1978) study, subjects rated pairs of faces on a 7 point scale for similarity/difference. Harmon (1973) based an early computer-driven face recognition system on ratings of face descriptors. Usually, ratings of similarity are made globally: that is, subjects are asked to rate faces on a single scale, ranging from (for example) 'not at all similar' to 'very similar'. Alternative conceptualizations and operationalizations are relatively unexplored. Researchers have not investigated whether 'highly similar' and 'easily mistakable' are coterminous, or correlated, nor have they systematically examined the dimensions governing similarity judgements. The psychometric properties of these similarity ratings are also very rarely reported, and there are some indications that this is an important failure: Lindsay (1994), for example, reports that facial similarity judgements show great inter-subject variability - an array of faces which appear highly similar to one observer may not appear at all similar to another observer.

This type of approach has the advantage of retaining a hold on the cognitive aspect of facial similarity: what is important, after all, for most face recognition research is *perceived* similarity. Rating studies obtain a 'direct' measure of this perceived similarity (notwithstanding the unexplored psychometric problems). The chief drawbacks of this technique are the dependence on subject ratings, and the statistical ramifications of this dependence.[241]

Scaling techniques

Although the use of subject ratings ensures the connection of similarity measures to cognitive process, such ratings are typically only useful for a small set of comparisons. Several authors have recognised the need to formulate similarity measures for larger stimulus samples, and have utilised forms of scaling technique to this end. Hirschberg, Jones & Haggerty (1978) obtained similarity

[241] The most severe of these ramifications is that a sizeable sample of subjects will be required to obtain reliable estimates, given the high inter-rater variance reported by Lindsay (1994), and from results reported in this thesis (see Chapter 7).

ratings of all pairs of faces in a large sample,[242] and entered these into a multidimensional scaling analysis (MDS), incorporating individual differences into the analytic model. Since MDS generates a dimensional basis, spatial distance measures can be used as a measure of similarity. Other, cognate approaches include Rhodes' (1988) study, which used a 'tree sorting' algorithm. Subjects chose the most similar pair of faces in a large set, the next most similar pair, and so on. The pairs of similarity ratings were entered into an FxF matrix, analysed (via non-metric MDS) either as FxF (averaged over subjects), or as FxFxN, where individual differences are of interest. The algorithm creates a tree with branches, and the similarity between faces is computed according to the rank of the branches in between them. Young & Yamane (1992) took extensive measurements from each of a set of faces, found Euclidean distances for each face on these 'axes', and then submitted the distances in matrix form to MDS. Davies, Shepherd & Ellis (1979) measured facial similarity as an interim step in an application of (hierarchical) cluster analysis. Forty eight subjects were individually given a set of 100 black and white photographs, and asked to sort them into piles on the basis of physical similarity. Sortings were entered into a 100 X 100 matrix, the main diagonal representing the frequency with which individual photographs were sorted into identical piles by subjects. Matrix entries were treated as similarity measures.

I think that these approaches - in principle - present the most satisfactory solution in the literature to the problem of measuring facial similarity. The recognition that a similarity metric must be based on a representational scheme capable of simultaneously representing all faces in a set is particularly important. The further recognition that similarity must be conceptualised as inherently multidimensional is also significant. However, the schemes discussed here do not go far enough: the dimensions of the representational space are implicit, and it is not clear that they can generate faces that are not in the set submitted to MDS in the first place. At any rate, the issue is not broached in these studies.

An alternate approach, which does not use MDS, but retains the notion of a multidimensional representational scheme, is the use of principal component analysis (PCA) exemplified in studies by Sirovich & Kirby (1987), O'Toole et al. (1994), and Craw & Cameron (1991). This approach appears capable of providing a set of generating dimensions that can accurately represent faces which were not included in the initial PCA. I will later argue that this is the most fruitful procedure to follow in quest of a facial similarity metric.

[242] The problem with subject ratings of large samples is that it is usually not feasible to obtain comparative ratings for each possible combination of faces. A sample of 100 faces, for example, sustains more than 9900 pairwise ratings.

Dimensions for the space

The multi-dimensional representational models outlined above do not adequately address the nature of the generating dimensions. In Valentine's several writings on the subject (1991a, 1991b, Valentine & Endo, 1992) he declines to suggest what these dimensions could be, except to note that they could be any aspect which discriminates between faces. Although Valentine may be correct in a theoretical sense,[243] the dimensions cannot be arbitrary if the representational model is to be taken as a model of perception and cognition. The nub of the problem is that the dimensions for a cognitive representational basis are probably tied up with primary visual processing, and it is extraordinarily difficult to make theoretical descriptions at this level commensurable with theoretical descriptions at higher cognitive levels. We encountered this problem earlier in the chapter, namely in the face recognition models of Bruce & Young (1986 - see page 106), and of Burton & Bruce (1990 - see page 109). In both models, the 'visual front-end' of the models was left almost completely unspecified.

The failure to address the nature of the dimensions significantly impairs the cognitive validity of multi-dimensional models. The measure of facial similarity I favour and use in this research makes explicit the nature of such dimensions, but I cannot claim that their nature lends the measure any direct cognitive validity. One of the goals of the research is to examine the correspondence between estimates of facial similarity derived from the measure, and estimates derived from direct *perceptions* of facial similarity, but this is not intended as cognitive validation of the dimensions. The intention is a practical one: if the measure has satisfactory measurement properties, and if it correlates reasonably with perceived similarity, then it may be a useful measure of facial similarity.

Although few authors have explicitly addressed the nature of the dimensions that might generate a multidimensional representational model, it is possible to identify dimensions that have been used in an implicit manner. These dimensions are typically embedded in 'knowledge systems' designed to facilitate accurate retrieval of faces, or in predictive statistical models. Some discussion of these is provided immediately below.

Descriptions as dimensions

Faces are undoubtedly visual stimuli, and a satisfactory account of face recognition processes must surely incorporate a notion of visual search. It is notable, however, that artificial recognition 'devices' can search a population of faces fairly accurately on the basis of verbal descriptions of

[243] If the goal is merely to provide a model capable of representing a set of faces, the dimensions do not matter. However, if the goal is to suggest dimensions that correspond in some degree to what humans use, then the nature of the dimensions is very important.

faces. In these recognition systems, a population of faces is represented in a 'description space': that is, verbal descriptions constitute the dimensions of the space.

One of the earliest demonstrations of such an approach is to be found in a seminal article by Harmon (1973). A large set of photographs of faces was rated on descriptive dimensions by a number of independent judges. Ratings were averaged across raters, and each face was coded as an n-tuple of mean ratings. Harmon showed that one could search this 'description space' by obtaining a rating on each of the dimensions, and by finding the nearest neighbour in the space to this 'target face'.[244] In this way, a subject could retrieve a face from the system by simply providing a description, or a set of ratings.

A similar, perhaps more substantial recognition system is reported by Ellis, Shepherd, Shepherd, Flin & Davies (1989). This system, known as FRAME, was developed in collaboration with the British Home Office. Faces are described on a number of dimensions in FRAME, some qualitative (judges' ratings), and some quantitative (anthropometric measurements). Witnesses provide a verbal description, which an operator codes, and a search is conducted, yielding the *k* most similar faces. Witness are free to identify a particular face, or to ask for the search to be moderated (by using the most similar face as the new 'key', or by redefining the search on specific features found on some of the *k* faces). The overall dimensions used in this system are: *Overall shape of face; Complexion; Hair; Forehead; Eyebrows; Eyes; Ears; Nose; Mouth; Chin; Facial hair; Physical peculiarities; Accessories.* Ratings on several sub-dimensions are also obtained. FRAME is shown in several studies to be effective: in particular, it is better than the usual 'mugshot album' searches on typical faces, but not on distinctive faces, where performance is roughly equivalent. The usefulness of this type of approach depends not so much on the use of descriptions, I suggest, as on the multi-dimensional representational scheme. Indeed, the use of descriptions as dimensions is greatly limiting, since each face needs to be rated by a sizeable number of judges, and the verbal nature of the descriptions must surely limit effectiveness.

Physical measurements as dimensions

A somewhat different approach to the problem of facial dimensions is the use of physical measurements. Burton, Bruce & Dench (1993), for example, investigated the ability of direct facial measurements to discriminate between male and female faces. Distances of various physical 'features' were taken from photographs of faces, on straight line segments (therefore capturing only

[244] In fact, Harmon suggested several search strategies (e.g. binary searches, selection of probable subsets) and several measures (Euclidean, correlations).

two dimensional information). A discriminant function analysis (DFA) was able to correctly classify about 85% of faces, in comparison to a baseline performance of human subjects at about 96%. But misclassifications bore little resemblance to faces humans mistook, and Burton et al. concluded that the DFA did not constitute a good psychological model. A similar effort, with three dimensional measures derived from a laser scanning device, gave about the same performance as the 2D DFA, but with fewer variables (6 instead of 12). When the 3D measures were combined with 2D measures, DFA results approximated that of human observers.

Few attempts have been made to generalise this technique to the identification of individual faces, and where it has, the physical measurements have typically been combined with other information (see the discussion of the FRAME system, above).

Eigenvectors of intensity maps in the picture plane

The most elegant solution to the 'dimensions problem', however, may lie in principal components analysis, or equivalently, in the auto-associative learning networks first described by Kohonen (1984). This approach has considerable practical advantages over those discussed above, and provides an appealing mathematical solution to the task of representing populations of faces, and not merely fixed sets. Since this is precisely the method to be explored empirically in later chapters, I will take considerable care to describe it accurately here.

A face can be represented and recognised with considerable accuracy in two dimensions, as for example in a photograph. This is partly because photographs retain information about the three dimensional form of faces (see Bruce, Hanna, Dench, Healey & Burton, 1992), and perhaps because the visual system is inherently adapted to the task of extracting three dimensional information from two dimensional retinal images. All of the information used by the visual system for face recognition purposes may be contained in projections of faces onto two dimensional surfaces like photographic emulsion. If we find a representational basis for these projections, it may serve as a heuristic cognitive model, and may also provide a metric for representing and comparing faces. The first of these concerns is not of immediate relevance to the thesis, but the second is.

The starting point of the PCA approach[245] is to conceptualize a digitized[246] face image as a two dimensional A=MxN array of intensity values.[247] An ensemble of images maps to a collection of

---

[245] The account offered here is a synthesis of the approaches outlined at some length by Turk & Pentland (1991), Sirovich & Kirby (1987), and Kirby & Sirovich (1990).

[246] Kirby & Sirovich (1990) offer separate accounts for analogue and digital images; the differences are not especially significant.

[247] At the level of the analysis, this MxN array is concatenated with respect to rows, and treated as an RX1 vector (where R = MxN).

points in this MxN space. Since face images will bear considerable resemblance to each other, this space will be relatively low-dimensional. PCA finds the vectors that generate this subspace.

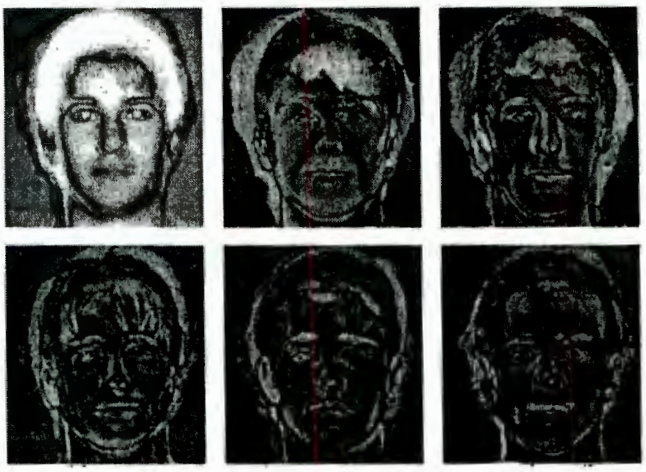
The ensemble of faces is $\Gamma_1, \Gamma_2, \Gamma_3, ..., \Gamma_M$. The average face of the set is $\Psi = \frac{1}{m}\sum_{n=1}^{M}\Gamma_n$ Each face differs from the average by the vector $\Phi_i = \Gamma_i - \Psi$. PCA requires that the face covariance matrix be formed, but the dimensionality of the matrix is large,[248] and the task of performing PCA on it is probably intractable (Sirovich & Kirby, 1987). However, for almost any particular set of face images, there will be fewer data points in the image space than the dimension of the space; there will only be k-1 meaningful eigenvectors rather than MxN of these (where k = number of images in the set). The task is therefore to first solve for eigenvectors of the MxN matrix, and then to take appropriate linear combinations of the images Φ_i. This technique reduces to a) constructing the MxN matrix, $L = A^TA$, where $L_{mn} = \Phi_m{}^T\Phi_n$, and ii) finding the M eigenvectors, v_i, of L (A = the face matrix). These determine the linear combinations of the M faces to form the 'eigenfaces' u_i, $\mathbf{u}_i = \sum_{k=1}^{M} v_{lk}\Phi_k$.

Each face in the set of face images can then be represented as a set of co-ordinates on these eigenfaces, or axes: the face will be perfectly reconstructed as a sum of the co-ordinate-weighted eigenfaces, provided that all the eigenfaces are used. If only a subset of the eigenfaces is used, the face will be imperfectly reconstructed, although this reconstruction may still be very accurate. By way of example, Figures 5.7 and 5.8 present an ensemble of six faces, and the six eigenfaces derived from an application of this technique.

For further details regarding the images, see Chapter 7, page 161 onwards.

Figure 5.7 Grey scale images of faces used in experiment 1a

[248] Sirovich & Kirby (1987) suggest that the dimensionality will be MxN.



For further details regarding the principal component analysis, see Chapter 7, page 161 onwards.

Figure 5.8 Eigenfaces produced by PCA in experiment 1a.

The benefits of this approach to the task at hand are considerable. Four are worth singling out for discussion.

The eigenfaces generated by the PCA allow the representation of the images in terms of a common set of reference axes (the eigenfaces). Individual facial images are linear combinations of these eigenfaces. This is a direct implementation of the type of multidimensional model discussed earlier in the chapter. It also provides a solution to the problem of identifying the dimensions for such a model: the dimensions are just the eigenfaces identified by the PCA. Of course, there is no evidence that 'eigenfaces' are used in human face recognition, and it is perhaps unlikely that such a complicated mathematical procedure could have a cognitive analogue.[249]

The multidimensional space generated by the eigenfaces has the associated advantage that well developed measures of spatial distance can immediately be used to determine nearest neighbours, relative density around particular points, and a variety of other useful indices. I suggest that the similarity of two images in the space can be determined by calculating a suitable distance measure, or by the dot product of the vectors representing the faces in the space. To some extent the identification of a suitable similarity measure is an empirical question, though, as one of the intentions is to develop a measure that corresponds reasonably closely to human perception of facial similarity.

If we admit some acceptable degree of error into the representation of faces in the space, we can represent the set of faces with considerably fewer eigenfaces than there are faces in the set. Thus,

[249] This is arguable, though, since O'Toole & Thompson (1993) show that the PCA approach is mathematically equivalent in most respects to an auto-associative neural network model of face recognition.

Cameron & Craw (1991) argue that about 40 eigenfaces is adequate to represent 100 faces, with about 5% error, and Kirby & Sirovich (1990) argue likewise. The amount of error that is acceptable here depends on the use to which the technique is to be put: O'Toole, Deffenbacher, Valentin & Abdi (1994), for example, argue that higher eigenfaces are important to rated distinctiveness and typicality of faces in the space. The relation of higher and lower dimensions to rated similarity is, again, an empirical matter, and will be explored in the research reported in later chapters.

One of the most useful spin-offs of the PCA approach is the ability to represent faces which were not in the original set, with a measurable degree of error. A new face, **F** is projected into the space by the matrix product, **uF**, and the result is the vector of eigenface co-ordinates of the new face. Indeed, this projection is recommended by Sirovich & Kirby (1987) as the first step in measuring the representation 'error' inherent in a lower dimensional space: the projection is used to reconstruct the face, and this reconstruction is measured against the original. Initial results from Sirovich & Kirby (1987) and from Craw & Cameron (1991) suggest that such error is typically very low, and that a fairly small number of eigenfaces may be able to adequately represent an entire population of faces.

Attractive as this approach appears, there are also several problems. In the first place, the approach unreasonably assumes the structural equivalence of images. Prior to analysis, images must be standardized, so that eyes in one image correspond in location to the eyes of other images - there would be no point in averaging ears and eyes, for example.[250] But faces are not structurally equivalent, and it is not possible to completely standardize the images prior to analysis by PCA. Figure 5.9 shows the 'blurring' that averaging incurs through the inevitable failure to completely standardize images.

[250] This objection is put more formally, and forcefully, by Craw & Cameron (1991), who point out that PCA assumes linearity, and that the axioms of linear systems demand that combinations of vectors in a linear subspace are again members of the subspace. In other words, the average image in a set of images must be a face, but this will not be the case unless images are structurally equivalent. An equally firm objection, apropos of another field of research, involving averages of pixel maps of faces, is lodged by Pittenger (1991), who points out that averages of three dimensional vectors are not equivalent to averages of the two dimensional projections of those vectors (but see a reply by Langlois & Roggman, 1991).

The standardization of images is not only difficult to accomplish with respect to location in the raw image space, but also with respect to aspects such as facial expression, ambient lighting, lighting of the face, etc. However, the variation across faces on these latter variables would not be nearly as great as the variation between the standardized images themselves, and Sirovich and Kirby (1987) have shown that several of these variables do not severely affect matters.

The image is averaged over pixel values of six faces

Figure 5.9 'Average' face, for face images shown in Figure 5.7.

In practice, images are aligned so that the pupils match (i.e. the left and right pupils occupy the same spatial locations on each image[251]). This ensures a close match of most faces. Craw & Cameron (1991) have outlined an alternate approach to the problem, which uses elements of the caricaturing method developed by Benson & Perrett (1991a, 1991b). That is, fiducial points are defined for a 'standard' image, and triangular tesselations are created by joining certain of these points.[252] Each face is then mapped onto this image, using bilinear interpolation, prior to the PCA.

In a recent unpublished paper, Hancock, Burton & Bruce (1994) used both methods (i.e. PCA on 'shape free' images, and PCA on 'shaped' images), and found a slight advantage for a combination of the 'shape free' and 'shaped' methods in predicting context free familiarity. Differences were slight, however, and there appears to be little advantage in the modified approach formulated by Craw & Cameron (1991).

From a practical point of view, standardizing images is laborious (much more so in transforming images to 'shape free' form), albeit necessary. An algorithm to automate the standardization would be a useful addition: several moderately successful attempts have been reported (Bowns & Morgan, 1993; Liu, Cheng, & Yang, 1993).

Conclusion

This chapter has surveyed a great deal of face recognition research, but the intention was not to provide a comprehensive review or evaluation. The aim was to raid this literature for a promising measure of facial similarity, and I suggest that the measure outlined in the final part of the discussion

[251] This involves scaling the images so that the inter-ocular distance is equal across all faces. An alternative technique is outlined by O'Toole & Thompson (1993).

[252] In practice, points are chosen on an *a priori* basis to represent the most important facial features (e.g. pupils, eyebrows, lips, etc), and the standard image is found by finding the mean location of these points across the set of faces.

is particularly promising. However, this thesis is in large part an empirical treatise, and later chapters will put the measure to the test.

Despite this avowed pragmatism, several issues which were identified in the review of the face recognition literature are worth referring back to here. Each of these suggests that the PCA approach to measuring facial similarity may shed some light on more general theoretical issues.

The dispute about whether face perception and recognition are featural or configural in nature appears to have resolved in favour of the configural view (at least in large part). More important for present purposes, though, is the finding that configural information depends on the general nature of the category (see page 102): since the relevant category is the human face, we need to know something about populations of faces in order to understand how the information could be used.

The formal face recognition models developed by Bruce and colleagues (see pages 104 and 108 onwards) are a welcome advance over the atheoretical empiricism of the field in earlier decades, but both leave the notion of a 'visual front end' unexplored. More recently, Bruce (1994) has acknowledged the importance of understanding the statistical properties of images, and also that PCA may be helpful. In particular, the 'eigenfaces' of a PCA analysis may constitute the canonical code postulated in the Bruce & Young model as the basis for the operation of Face Recognition Units.

One of the key propositions in Valentine's multidimensional model of face representation and recognition is that representations of out-group (or simply, less familiar) faces are 'less developed' than representations of in-group (or familiar) faces. The model does not detail how such a state of affairs could come about, but PCA (and particularly, its neural network formulation) provides an explanation. Components capture the statistical variation of faces, and groups of faces that are under-represented in the face space will be represented by fewer components (hence, less accurately).

These are incentives for further research rather than pointers to later chapters. In the rest of the thesis - that is, after Chapter 6 - I will concentrate on the validation of a PCA-based measure of facial similarity, and the application of such a measure to identification parades.

Chapter 6

Statistical Inference on Measures of Lineup Fairness

Introduction

In Chapter 4, I reviewed research on identification parades and showed that psychologists have been active. In particular, legal psychologists have contributed an especially valuable corpus of research which concerns the development and evaluation of measures of lineup fairness. Measures proposed thus far are usually of a descriptive nature, and little has been written about their inferential use. In this chapter I investigate some commonly used measures, and suggest ways in which inferential statistical considerations may improve their usefulness. My approach will be to trace the development of measures of fairness chronologically, and to raise statistical considerations in relation to each measure. I argue that it is important to develop ways of reasoning inferentially about the measures, and I show a few consequences of not doing so. As I reviewed these measures in an earlier chapter, some of the descriptions here are repetitions of the earlier material. In particular, many of the tables in the earlier chapter re-appear here to add clarity to the discussion.

Departure from expected values

In Chapter 4 I noted that recent psychological interest in measures of lineup fairness appears to stem from an experiment conducted by Doob & Kirshenbaum (1973). The experiment tested a police lineup used in a Canadian case for fairness. In the part of the experiment which is of interest in this chapter, Doob & Kirshenbaum showed a photograph of the identification parade to 23 'mock witnesses' (witnesses who had not been present at the original crime), along with the 'description' of the suspect given to police by the witness. They reasoned that subjects who had not been present at the crime should not be able to identify the suspect with a probability exceeding that expected on the basis of random selection (1/number of lineup members = 1/12 = 0.083). 64% of the witnesses correctly identified the suspect. Doob & Kirshenbaum reported that the probability of this occurring is less than 0.001, and concluded that the lineup was biased.

Doob & Kirshenbaum suggested that this method - which has come to be known as the method of the mock witness - could be used in general to assess the fairness of identification parades. The method posits that a lineup is fair when the proportion of witnesses choosing the suspect

does not exceed that expected on the basis of mere random choice. Doob & Kirshenbaum report a probability value to support the conclusion they make in their study, but they do not recommend a general method for determining when the proportion exceeds that expected on the basis of random choice. I suggest that the following conceptualization provides a suitable probability model of random witness choice, and can be used to evaluate the proportion of suspect identifications against a suitable null hypothesis.

> There are N mock witnesses, and there are k lineup members. The probability that a witness will choose the suspect, given that the choice is made randomly, will be 1/k. We assume that witnesses choose independently of each other. Then each individual witness choice can be thought of as a Bernoulli trial, with probability of success = 1/k. The number of trials, q, in which a successful choice is made will take the Binomial probability distribution, i.e. the number of correct identifications will be distributed as (1) below

$$p_Q(q) = \binom{N}{q}\left(\frac{1}{k}\right)^q\left(\frac{k-1}{k}\right)^{N-q} \tag{1}$$

The cumulative distribution of q will take the form of the cumulative binomial distribution, i.e. (2) below

$$F_Q(t) = \sum_{q \le t}\binom{N}{q}\left(\frac{1}{k}\right)^q\left(\frac{k-1}{k}\right)^{N-q} \tag{2}$$

The cumulative distribution gives the probability that 1 witness, or 2 witnesses, or 3, ..., or t identified the suspect just by chance.

For Doob & Kirshenbaum's experiment, where 11 out of 21 mock witnesses identified the suspect, the exact probability that this occurred is 1.96×10^{-7}, and the probability that 11 or more of the mock witnesses identified the suspect is 2.15×10^{-7}. This value is very small, and indicates the low likelihood of 11 correct random choices.

The calculated values in (1) and (2) can, of course, always be compared to a conventionally agreed upon 'level of significance', but the exact probability is more informative. Both distributions are available as functions in many commercially available software packages.

In Table 6.1, frequencies are reported for several hypothetical lineup configurations, and probabilities calculated under (1) clearly reflect the bias against the suspect in each.

	Lineup Member						
	1	2	3	4*	5	6	Probability
a	6	4	7	30	9	4	3.62×10^{-9}
b	8	9	8	10	9	6	0.12
c	1	3	4	10	1	1	0.0005

* = suspect

Table 6.1 Suspect identification probabilities in a number of hypothetical lineups.[253]

There are other methods of evaluating the number of accurate identifications made by mock witnesses. Malpass & Devine (1983), and Buckhout, Rabinowitz, Alfonso, Kanellis & Anderson (1988), for example, use a one-sample z-test to compare the observed proportion of correct witness choices to that expected under an equiprobability model.[254] It is a simple enough task, but it relies on the fact that the normal distribution is the limiting distribution of the binomial: the approximation is good for large samples, but may not be very good for small samples, especially when the parameter 1/k is small. In mock witness experiments one, or both, of these conditions may not be true: Doob & Kirshenbaum, for example, used 21 subjects, and there were 12 lineup members (i.e. 1/k = 0.083). I suggest that, in most cases, it is better to use (1) above to evaluate the rate at which mock witnesses identify the suspect. It yields an exact probability estimate, and the underlying probability model accurately represents the nature of the mock witness task.

But there is a more general disagreement with the z-test comparison, which applies equally to the method suggested in (1) above. Wells, Leippe & Ostrom (1979) argue that the z-test of equivalence used by some researchers to compare the proportions of identifications suffers from a dependency on sample size. In order to conclude that a lineup is biased, researchers would simply need to obtain a sufficiently large sample and conduct a mock witness evaluation of the lineup. Consequently, we need to view claims of bias supported by mock witness tasks with extreme caution. This argument is correct, but it applies to almost all instances of significance testing, and not merely to significance test evaluation of lineups. The appropriate use of the mock witness task is dependent on the good judgement of the researchers who apply it, and this is most certainly true of any inferential statistical procedure.

The model outlined in (1) accurately represents a lineup task in which witnesses are choosing randomly. For this reason, I argue that it is a suitable model for the null distribution in the mock

[253] Data in this table are presented under the assumption that the 'not present' option was unavailable in the lineup tasks. This does not materially alter the calculation of the probability.

[254] Several other methods are possible: one equivalent would be to treat a standardized residual determined from a Chi-square goodness-of-fit procedure as a normal deviate.

witness task. It is unlikely, however, that it will accurately represent choosing patterns in particular mock witness tasks, or in real situations, since in both cases witnesses possess specific information about the identity of the perpetrator, and will not choose randomly. Navon (1990a, 1990b), apropos of a related issue, argues for a conceptualization that views identification decisions as conditional on the similarity of foils to information possessed by witnesses regarding the identity of the suspect. He argues that the correct method is to determine how likely a random match would be of the perpetrator and an innocent person whose features resemble those of the perpetrator. This probability can be expressed as $p(ids|r_{st} > r_0$ and $N=n)$.[255] Its successful calculation requires estimation of population parameters of physical similarity, and it reduces to the expression given by (1) once the size of the sub-population that meets the similarity requirement is known. This is an interesting approach, but I am not in complete agreement with it: in particular, I think that the task of setting a cut-off point for physical resemblance invites vagaries, and fails to take into account that even a small degree of resemblance may influence the probability of an identification.

Functional and Nominal size

The intention of Doob & Kirshenbaum's study was to provide a measure of lineup fairness. Information regarding lineup fairness is provided to some extent by the proportion of accurate mock witness choices, but it is certainly not complete. An important additional feature of a lineup appears to be the number of plausible foils that it contains.

Wells, Leippe & Ostrom (1979) coined the term 'functional size' to deal with this type of situation: functional size is intended to provide an index of the number of plausible lineup members, and is therefore also a measure of lineup fairness. The measure relies on the mock witness task introduced by Doob & Kirshenbaum, but avoids certain problems encountered by Doob & Kirshenbaum's procedure for evaluating fairness when 'null' and 'perfect' foils are present in the lineup. These problems were discussed in Chapter 4 (page 66 onwards). Wells et al. suggest an alternate measure, which attempts to avoid some of the problems. The measure is defined as D/N (where D is the number of mock witnesses who choose the defendant, and N is the total number of mock witnesses), and represents the proportion of mock witnesses who identify the suspect. It is insensitive to the problem of null and perfect foils, and Wells et al. suggest that the measure has a useful interpretation when transformed to its reciprocal (N/D), which is the number of functional members of a lineup, hence the term 'functional size'. It is this index that they suggest as a measure of lineup fairness. A

[255] Ids = identification; r_{st} is the resemblance between the subject and the target (perpetrator), r_0 is the lower bound of the similarity that can be inferred from the lineup, and n is the size of the population with members who meet the resemblance criterion.

lineup is fair when the 'functional size' and the 'nominal size' are identical. In the hypothetical lineups d and e of Table 6.2,[256] which represent lineups with clearly divergent numbers of plausible foils, functional size is 2 and 6 respectively, and this difference corresponds quite well to the apparent difference in fairness of the lineups.

	Lineup Member							
	1	2	3	4*	5	6	Not Present	Functional Size
d	5	3	6	30	9	3	4	2
e	8	9	8	10	9	9	7	6
f	1	3	20	10	21	4	1	6

* = suspect.

Table 6.2 Functional size in a number of hypothetical lineups.

Malpass (1981) has convincingly argued that the statistic suggested by Wells et al. does not provide an index of the 'size' of the lineup. The measure of 'functional size' depends only on the proportion of accurate mock witness identifications, and takes no account of the distribution of identifications across the foils. It is possible for the functional size of a lineup to be identical to its nominal size and for the distribution of identifications to exhibit a clearly different picture about the number of plausible foils. Lineup f of Table 6.2 is one example (the functional size of the lineup is 6, suggesting 6 plausible lineup members),[257] and it is easy to imagine many others. Indeed, the only type of situation in which *D/N* will give a reasonable indication of the number of plausible foils may be when the distribution of identifications is nearly uniform, as is the case in lineup e of Table 6.2.

Nevertheless, Wells et al.'s measure of functional size is useful under certain circumstances, and is quite widely used in lineup research. It is worth considering how to reason inferentially about it. The measure is identical in an important respect to that suggested by Doob & Kirshenbaum. Doob & Kirshenbaum think that the difference between the observed and expected proportions of accurate mock witness identifications is the critical index, but Wells et al. think that the critical index is simply the observed proportion of accurate identifications (and its reciprocal). The statistical considerations outlined in (1) above therefore also apply to the measure advanced by Wells et al. In particular, I suggest that the most appropriate way to apply inferential reasoning here is to express the observed proportion of accurate identifications as a confidence interval, and then to take reciprocals of the

[256] Note that this table also appears in Chapter 4: as indicated in the introduction to the present chapter, several of the tables of Chapter 4 are repeated here for ease of presentation and discussion.

[257] This particular distribution of lineup identifications is not as unlikely as it may seem. There are several reported instances of mock witness trials where foils have drawn more identifications than the suspect (cf. Buckhout, Rabinowitz, Alfonso, Kanellis & Anderson, 1988).

endpoints in order to express the interval in terms of so-called functional size. I will discuss two methods of determining a confidence interval. The first method, (3) below, is well known, relies on approximation theorems, and is appropriate for large samples.

$$\frac{D}{N} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{D(N-D)}{N^3}} \tag{3}$$

Z is a standard normal deviate chosen according to the desired level of confidence, α, and the other terms are as defined in the discussion above.[258]

The second method, (4) below,[259] improves the approximation at the cost of some computational effort.

$$\frac{N}{N+Z^2}\left[\frac{D}{N}+\frac{Z^2}{2N} \pm Z\sqrt{\frac{\frac{D}{N}\left(1-\frac{D}{N}\right)}{N}+\frac{Z^2}{4N^2}}\right] \tag{4}$$

Equation (4) gives an approximate 100(1-α) percent confidence interval for D/N. An interval for functional size can be obtained in turn by taking the reciprocals of the endpoints.[260]

In lineup e of Table 6.2, 0.17 of the mock witnesses identified the defendant (i.e. functional size is 6), and the 95% confidence intervals around this value calculated under (3) and (4) above are (0.07; 0.26), and (0.10; 0.28) respectively. Taking the reciprocals of the endpoints gives intervals of (3.9; 14.3) and (3.57; 10). A larger sample would have narrowed these interval estimates considerably: indeed, the upper endpoints are impossible outcomes! (This is a failure which is attributable not to the interval estimation, but to the measure itself. Whenever a suspect is chosen at rates lower than expectation, functional size will be greater than the number of lineup members). The range of the interval indicates the importance of reasoning inferentially about functional size. An estimate of functional size which is derived from a relatively small sample of mock witnesses should not, in my opinion, be accorded the same weight as one derived from a much larger sample. Confidence intervals incorporate just such a corrective weighting. In the absence of inferential reasoning of this kind, a particular estimate of functional size has uncertain meaning and must be of limited value.

---

[258] It is customary to distinguish population parameters and sample estimates of the parameters at the level of notation. I do not follow this convention on a systematic basis in this chapter; I do so only where it facilitates the clarity of the discussion. In most instances the usage I mean will be clear from the context.

[259] This correction is discussed by several authors; see Hays (1994), for example.

[260] Rick Gonzalez, of University of Washington, reviewed a draft of this chapter, and noted in correspondence that the transformation to reciprocals may result in biased intervals. He suggests an alternative technique, which is to determine the interval as $(\frac{1}{\hat{p}} - \varepsilon', \frac{1}{\hat{p}} + \varepsilon')$, where ε' is defined with respect to the standard error of $\frac{1}{\hat{p}}$. He has identified a potential estimator for this standard error, but this will not be detailed here. He agrees, however, that both the method he suggests, and that outlined in this chapter, will work increasingly well with increasing N.

Telling a jury that the functional size of lineup e in Table 6.2 is 6 without at the same time giving an indication of the error of estimate is not completely honest testimony.

Effective Size and Defendant Bias

Malpass (1981; Malpass & Devine, 1983), has provided a thorough analysis of the contributions made by Doob & Kirshenbaum and Wells et al. to the measurement of lineup fairness. In his analysis he argues for a distinction between *lineup size* and *lineup bias.* Lineup size refers to the number of plausible members that the lineup contains, and it contributes directly to the fairness of the lineup by decreasing the probability that the defendant is identified by a witness who wilfully chooses at random. Lineup bias, on the other hand, is the extent to which mock witnesses choose the defendant at rates greater (or smaller) than chance expectation. Because both these components contribute to the fairness of a lineup, a measure of each is required.

The first measure to be considered is what Malpass calls 'effective size', which is close in meaning to the measure of functional size. In order to evaluate a lineup, Malpass argues, the critical thing to know is how many plausible foils it contains. Although Wells' measure of functional size attempts to estimate this quantity, it has several faults. Malpass suggests an alternative measure to functional size, 'effective size':

$$\text{Effective size} = k_a - \sum_{i=1}^{k_a} \frac{|o_i - e_a|}{2e_a}$$

where o_i = the (observed) number of mock witnesses who choose lineup member i; e_a = the adjusted nominal chance expectation ($N*[1/k_a]$); k_a = the adjusted nominal number of alternatives in the lineup (original number - number of null foils).[261]

The intent of the measure is to reduce the size of the lineup from a (corrected) nominal starting value by the degree to which members are, in sum, chosen below the level of chance expectation. As is clear from the formula, the absolute value of the difference between observed and expected values is taken,[262] and the sum is divided by 2, in order that the calculation reflect the sum of differences where $o_i - e_i < 0$. (Malpass reasons that the important departures are those below chance expectation, since lineup members who fail to draw identifications are poor foils).

The idea behind the measure of effective size is interesting, and importantly, it attempts to use information regarding the distribution of identifications across the lineup, which is necessary if we are

[261] The notation used here differs slightly from that given by Malpass (1981), in order to remain consistent with notation used earlier in this chapter.

[262] Since the sum of signed first order differences is zero. Squaring the differences is an alternate method.

to say something about the overall constitution of the lineup. However, there are several features which seem a bit problematic.

The first is the assumption that null foils are totally implausible - indeed, non-members of the lineup. This is not the case. Null foils have a positive probability of occurring, especially when the number of lineup members is large, and the sample of mock witnesses relatively small. In Doob & Kirshenbaum's experiment, for example, there were 12 lineup members, and 21 witnesses. From (1), we can show that the probability that a particular lineup member will be a null foil is[263]

$$\binom{21}{0}\left(\tfrac{1}{12}\right)^0\left(\tfrac{11}{12}\right)^{21} = \left(\tfrac{11}{12}\right)^{21} = 0.16$$

The probability that there will be exactly one null foil (which could be any of those constituting the lineup) should also be considered, and can be calculated with repeated applications of the binomial theorem

$$12\left(\tfrac{11}{12}\right)^{21}\prod_{i=2}^{12}\left(1-\left(\tfrac{11}{12}\right)^{21}\right) = 0.28$$

Finally, the probability that there will be at least one null foil should be determined for sake of completing the argument. To do this we calculate the probability that there are 1, or 2, or ..., j, ..., or 12, null foils in a 12 member lineup:

$$\sum_{j=1}^{12}\left[j\left(\tfrac{11}{12}\right)^{21}\prod_{i=j}^{12}\left(1-\left(\tfrac{11}{12}\right)^{21}\right)\right] = 0.31$$

Clearly, the assumption that a null foil warrants discarding from a calculation of effective size must be viewed with some scepticism. Null foils have a good chance of appearing in distributions of identifications obtained with the mock witness task. The chances are smaller when larger samples of mock witnesses are used, but are only negligible when these are very large. For example, if we assume that Doob & Kirshenbaum used 50 subjects, the probability of obtaining at least one null foil shrinks to 0.12, but to reduce this to say, 0.01 would require a sample closer to 100 subjects. Null foils should not simply be dismissed from calculations of lineup size since they may occur with appreciable probability. I suggest revising the definition of effective size to include null foils, i.e.

$$\text{Effective size} = k - \sum_{i=1}^{k}\frac{|o_i - e|}{2e} \qquad (5)$$

[263] This calculation assumes for convenience that choosing a particular foil consitutes a successful trial, and that choosing any other member of the lineup constitutes a failure.

The assumption underlying the notion of effective size is an appealing one. One or more of the foils in a lineup may present an inadequate test of a witness who has little more than only very general knowledge of the appearance of the offender, and we shouldn't take the ability of a witness to reject such foils very seriously. The calculation of effective size acts on the assumption by reducing the nominal size of the lineup as a function of the departure of proportionate identification of individual foils from that expected by an equiprobability model.

For many distributions of identifications the measure does seem to give an indication of the number of foils that could reasonably be considered present. Lineups g, h, and i in Table 6.2 are clear examples.

	Lineup Member							
	1	2	3	4*	5	6	E_a	E_b
g	0	25	5	25	3	2	2.83	3.00
h	10	10	9	10	11	10	5.90	5.90
i	1	0	12	12	0	11	3.17	3.17
j	7	7	7	24	8	7	4.60	4.60
k	12	6	9	13	14	6	5.10	5.10
l	6	19	3	20	8	10	4.45	4.45

E_a = effective size calculated with adjustment for null foils.
E_b = effective size calculated without adjustment for null foils.
* = suspect

Table 6.3 Effective size in a number of hypothetical lineups.

However, the measure also produces estimates that seem intuitively at odds with visual inspection of particular lineup distributions. In lineup j of Table 6.2, for example, effective size is 4.6. It is difficult to see what this means: clearly, foils are drawing identifications at a rate close to the expected value, it is just that one lineup member is drawing far more identifications than expected (indeed, is chosen by 40% of the witnesses). If this foil happens to be the suspect, it is difficult to see that a conjectured effective size of 4.6 has the meaning that is intended: a mock witness, armed with only superficial knowledge of the physical identity of the perpetrator, does not have a 0.22 (i.e. 1/effective size) chance of identifying the suspect, but indeed, is more likely to succeed closer to half the time. The foils are not plausible, even though they are chosen at a level fairly close to expectation. It is possible, of course, to argue - in defence of the measure - that lineup size and lineup bias are conceptually distinct and that bias can be ascertained separately. But it is then difficult to understand what meaning the notion of effective size has, because it doesn't seem to estimate the number of plausible foils.

In particular, when there are several near-null foils, and there is no bias against the suspect, the measure gives a reasonable indication of the effective number of lineup members, but when the

distribution is less uniform, the measure is difficult to interpret. Lineups j, k and l in Table 6.2 show this to be the case: it is difficult in each case to see that the calculation gives a good indication of the plausible number of foils.

The problem may be that the notion of 'effective size' inherently suggests that only a discrete number of foils are plausible alternates to the suspect - there can hardly be exactly 3.2 plausible foils in a five member lineup! The question of plausibility must surely be one of degree. Particular effective size estimates only seem really convincing in cases where i) a nearly uniform distribution of mock witness choices exists, or ii) certain foils are chosen at near-zero rates. Although the structure of the lineup appears to invite discrete-scale reasoning, it may be fruitful to think in terms of continuous variation of plausibility. In this sense the proportion of mock witness choices that each lineup member attracts serves as the best estimate of plausibility.

For this reason I suggest that researchers using the mock witness task should routinely consider the administration of a goodness of fit test, where the observed proportions are evaluated against the proportions expected under an assumption of equiprobability. This would give an indication of the reliability of the departure of the distribution of lineup choices from uniformity. A Pearson χ^2 goodness of fit test on a hypothetical lineup is reported below in Table 6.4 for purposes of demonstration (a likelihood ratio χ^2 test could also be used).

	Lineup Member					
	1	2	3	4*	5	6
Observed freq.	3	10	17	23	9	22
Expected freq.	14	14	14	14	14	14
Residual	-11	-4	3	9	-5	8
Std. Residual	-3.21	-1.17	0.88	2.63	-1.46	2.33

$\chi^2 = 22.57$; df = 5; $p < .01$ * = suspect

Table 6.4 Hypothetical lineup demonstrating goodness-of-fit test

Standardized residuals presented in the last row can be used to assess the extent to which different foils are chosen at levels below (or above) expectation. These take an approximately normal distribution (the approximation is much better for large samples).

The goodness of fit test outlined above evaluates a complex null hypothesis. Any of a large number of possible departures from a uniform distribution of witness choices may lead to rejection of the hypothesis. Its utility is therefore limited. It may be more useful for eyewitness researchers to construct a confidence interval around estimates of effective size, notwithstanding the reservations regarding the interpretation of the measure, which I expressed earlier.

Unfortunately, both the measure of effective size suggested by Malpass (1981) and the modification to the measure suggested in (5) do not easily lend themselves to known inferential methods,[264] and I am unable to make suggestions particular to the measures.

There are alternate strategies. The rationale of the effective size measure is to reduce the nominal size of the lineup in proportion to the deviation of the observed frequencies from expectation. A related measure, with a similar rationale, is the "index of diversity" (I), and its standardized form, the "index of qualitative variation" (Q) (Agresti & Agresti, 1978). The measures compute the variation in a one dimensional array of frequencies.

$$\underline{I} = 1 - \sum_{i=1}^{k} \left(\frac{o_i}{N} \right)^2 \tag{6}$$

$$\underline{Q} = \frac{k}{k-1} \left[1 - \sum_{i=1}^{k} \left(\frac{o_i}{N} \right)^2 \right] \tag{7}$$

k, N and O_i are all as defined earlier.

I varies between 0 and $\frac{k-1}{k}$, and Q is standardized to vary between 0 and 1. The measures are at a minimum when one array item attracts all the available responses (i.e. all mock witnesses choose one lineup member), and at a maximum when responses are equally distributed over array items.

Agresti and Agresti (1978) have shown how the measures I and Q can be used inferentially, and I suggest methods whereby their results are used to find confidence intervals for estimates of the 'effective size' of lineups, and to test for differences between independent estimates of 'effective size'.

The variance of I can be estimated by

$$\sigma_I^2 = \frac{4}{N} \left[\sum_{i=1}^{k} \left(\frac{o_i}{N} \right)^3 - \left[\sum_{i=1}^{k} \left(\frac{o_i}{N} \right)^2 \right]^2 \right] \tag{8}$$

---

[264] In particular, the use of the modulus seems to render the measures intractable. It can be shown that the size of the measures will depend on the *number* of observed frequencies below expectation, i.e.

$$k - \sum_{i=1}^{k} \frac{|o_i - e|}{2e} = j + \frac{k}{N} \sum_{i=j+1}^{k} o_i$$

where j = the number of foils chosen at rates equal or below expectation, and other terms are as previously defined.

A large sample[265] 100(1-α) percent confidence interval for I can be estimated by

$$\underline{I} \pm Z_{\frac{\alpha}{2}}\, \sigma_I \tag{9}$$

Similarly, a large sample 100(1-α) percent confidence interval for the difference between two independent I's can be estimated by

$$(I_1 - I_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\sigma_1^2 + \sigma_2^2} \tag{10}$$

This interval can also be used to test the hypothesis that the values of I differ (the null hypothesis is rejected if the interval contains 0).

The notion of effective size rests on an interpretative strategy: deviation from a uniform distribution is interpreted to mean that fewer than k plausible foils constitute the lineup. Whereas Malpass' original formulation of effective size and the modified form set out in equation (5) attempt to estimate this directly from the departure of the distribution from uniformity, I suggest that a similar interpretation can be achieved by determining the values of I[266] that correspond to the presence of *m* foils in a lineup with *k* members.

Imagine that a k member lineup has 1, 2, ..., or *m* plausible members *(m ≤ k)*, and that only these members draw witness choices. Assume further that witness choices are equally distributed across the plausible members. Now we want to calculate values of I for each of the *k* lineup configurations. We can show that this will be the arithmetic sequence[267] $\left\{\frac{m-1}{m}\right\}_{m=1}^{+\infty}$ Table 6.5 shows the possible values of I for all possible numbers of plausible foils in lineups varying in nominal size between 1 and 12.

	Number of plausible foils											
	1	2	3	4	5	6	7	8	9	10	11	12
I	0	.50	.6667	.750	.80	.8333	.8571	.8750	.8889	.90	.9091	.9167

Lineups share initial sequences, so it is unnecessary to present values for different lineup sizes: to find the value of I corresponding to the number of plausible foils, for any size of lineup, simply read off the value. The series converges to 1 for large values of k. It is important to round to several digits for accuracy, increasingly so with larger values of k.

Table 6.5 Possible values of I for lineups varying in nominal size, and in number of plausible foils.

[265] Agresti et al. contend that the approximation will be good for even moderate sample sizes, provided none of the observed frequencies approaches N.

[266] I restrict myself here to the measure I, for sake of convenience. A similar argument can be made for Q.

[267] Under the restrictive assumptions, $\underline{I} = 1 - \sum_{i=1}^{m} \frac{1}{m^2} = 1 - \frac{1}{m} = \frac{m-1}{m}$.

Thus, a calculated value for $\underline{I}$ of 0.8 would be interpreted to mean that there are 5 plausible foils in the lineup used for the calculation. Interpretation can be made a little easier by transforming calculated values of $\underline{I}$ directly to 'effective size', i.e.

$$'E' = \frac{1}{1-\underline{I}} \qquad (11)^{268}$$

We can make this interpretation a little more rigorous, I suggest, by calculating confidence intervals around point estimates of $\underline{I}$ and transforming upper and lower bounds to reflect effective size. Table 6.6 presents point and interval estimates of $\underline{I}$ and effective size ('E') for a hypothetical lineup.

Lineup Member							
1	2	3	4*	5	I	'E'	
3	24	30	33	10	.7326	3.740	Point estimate
					.7015	3.503	Lower bound
					.7637	4.232	Upper bound

* = suspect

Table 6.6 Point and 95% confidence interval estimates of $\underline{I}$ and 'E' in a hypothetical lineup.

The confidence intervals provide probabilistic evidence for the proposition that there were fewer than 5 plausible lineup members. It is interesting to note that the estimate of effective size under equation (5) is 3.65, which is very close to the point estimate in Table 6.6, although this needs further exploration if we are to take 'E' seriously as an approximation of the formulations of effective size set out earlier. The application of probabilistic reasoning to measures of effective size is important, though, regardless of the precision of the approximation. In line with the argument made earlier apropos of functional size, I suggest that it may be misleading to simply present jurors (or any other interested party) with a point estimate of effective size, because mock witness identifications are inherently prone to sampling variation, and so, therefore, are measures of effective size. A confidence interval presents the evidence in a more satisfactory manner.

Defendant bias

Malpass (1981) points out that effective size does not provide a measure of bias towards the suspect, but an estimate of the number of plausible foils present in a lineup. Thus, it is possible for a suspect to participate in an unbiased lineup of very small effective size. Imagine that only three members

[268] Values of $\underline{I}$ in Table 5 are calculated by assuming a hypothetical 'effective size' i.e. $\underline{I} = \frac{m-1}{m}$. The reciprocal of this - given above as (11) - therefore transforms $\underline{I}$ back to hypothetical 'effective size'.

(including the suspect) of a ten member parade are plausible choices. If the suspect is chosen by mock witnesses with a probability of 1/3, there is no bias toward the suspect, despite the large number of redundant foils. We should not reject this lineup because of the differential probability of successfully choosing the suspect, but rather because of the increased risk that the suspect will be chosen at random from the lineup with three effective members (i.e. at a rate of 0.3̄ per identification).

Bias towards (or against) the suspect is another matter. Malpass advises that, in principle, we follow Doob & Kirshenbaum's method, which was described earlier. Here bias is equated with the departure of the proportion of suspect identifications from that expected under an assumption of equiprobability. He suggests that the 'size' of the lineup is better estimated by the calculation of 'effective size', which I discussed at some length immediately above. Thus, defendant bias will be measured by departures of the suspect identification probability from 1/[effective size of lineup]. The idea here is that the likelihood of being selected randomly by a witness is less a function of the mere size of the lineup than a function of the number of plausible foils present in the lineup.

I expressed reservations about the definition of effective size earlier, so I won't repeat them here. I simply pause to suggest that the re-definitions offered in (5), and (11) be used in calculations of defendant bias.

Percentage below expectation

Malpass & Devine (1983) also suggest a method for evaluating the suitability of individual foils, which is closely related in principle to the measure of effective size. The critical datum for evaluating the suitability of an individual foil, Malpass & Devine suggest, is the extent to which the foil is chosen below chance expectation in a mock witness task. Thus, if foil 1 is chosen from a ten member lineup with some low probability, this would suggest that the foil does not adequately represent the description given to the police by the eyewitness. (The question of the extent of departure from chance expectation is a thorny one. Malpass & Devine suggest leaving the decision to the fact finders).

An alternative approach to measuring parade fairness would then be to set a minimum size criterion, and to determine whether the parade meets the minimum size (the estimate of size would be determined by including only plausible foils - and the suspect - in the total). This approach would apparently be intuitively appealing to lawyers and judges.

I suggest here that decisions about whether an individual foil meets some minimum identification criterion should be made in terms of the probability model underlying the nature of the mock witness task, outlined in (1) above. This is an important consideration, since we can reasonably anticipate considerable departures from expected rates of identification, especially when relatively small

samples are used. For example, imagine that we use a lineup with 10 members, a sample of 30 mock witnesses, and a minimum criterion size of 67%. Then we can apply the cumulative binomial distribution outlined in (2) to determine the probability that any particular foil is chosen at a rate below 0.67 of chance expectation[269]:

$$\sum_{i=0}^{2}\binom{30}{i}\left(\tfrac{1}{10}\right)^{i}\left(\tfrac{9}{10}\right)^{30-i} = 0.41$$

This is a very high probability, and only a large increase in sample size will reduce it to an acceptably low level. A sample size of 100, for example, will reduce this probability to 0.01. But note that this is only the probability that *a particular* foil is chosen at a rate below 0.67 of expectation, and the probability that *at least one* foil is chosen at such a rate will be higher.

I therefore suggest that it would be inappropriate to simply disregard foils chosen at rates below some minimum criterion. A better method may be to construct confidence intervals around the observed proportion of identifications that each foil receives, and to apply the 'minimum criterion' test to the endpoint(s) of the intervals. This would have the benefit of attaching some level of probabilistic confidence to any decisions taken about the plausibility of foils. Methods for constructing intervals for both small and large sample cases are presented earlier in this chapter i.e. see (3) and (4) above. Since intervals would need to be constructed for each foil, it is worth considering some correction for a potential increase in the Type I error rate. A Bonferroni type correction, setting $\alpha' = \alpha/k$, is an obvious but conservative choice. Table 6.6 provides intervals for a hypothetical mock witness response to a lineup array, without correction for a potentially increased Type 1 error rate. Two members of the lineup have intervals that fall completely below the expected proportion of choices, and in the case of one of these the interval falls completely below even 50% of expectation. By the preceding argument we should conclude that two of the foils attract too few mock witness choices to be considered plausible.

[269] This calculation assumes for convenience that choosing a particular foil consitutes a successful trial, and that choosing any other member of the lineup constitutes a failure.

	Lineup Member				
	1	2	3	4*	5
Observed	3	24	30	33	10
Lower bound	0.00	0.16	0.21	0.24	0.04
Upper bound	0.06	0.32	0.39	0.42	0.16

* = suspect. Expected proportion is 0.20

Table 6.7 Interval estimates of foil identification proportions in a hypothetical lineup.

We could also have conducted a goodness of fit test, as demonstrated in Table 6.4 above, and used the standardized residuals to evaluate the extent to which foils were under-chosen by mock witnesses. Estimation of intervals seems a more satisfactory method, though, since it is closer to the original metric.

Lineup Size and Lineup Bias

Malpass & Devine (1983) argue for a conceptual distinction between 'lineup size' and 'lineup bias', and the measures they suggest reflect this distinction. However, they also argue that these measures will often be empirically dependent. If an implausible foil fails to attract identifications, the size of the lineup is automatically reduced. If these identifications are randomly distributed over the remaining lineup members bias will be unaffected, but if the identifications are disproportionately distributed, bias will be affected. They suggest that a similar argument can be made for the case where bias is increased: changes in size will depend on whether the other lineup members 'lose' identifications proportionately.

The case for the dependency of lineup size and lineup bias is compelling, but primarily because the measures are partly confounded in their definitions. A lineup is biased to the extent that observed frequencies depart from expectation (usually calculated only for suspect identifications), and a line-up's effective size departs from its nominal size to the extent that observed frequencies are lower than expectation. It follows that suspect identifications occurring below expected levels must lead to estimates of effective size that are lower than nominal size.

I have no clear answer to this problem. I think that it stems from two deeply rooted assumptions, and that the assumptions may need careful examination. In the first place, as I noted earlier, foil plausibility is generally assumed to be an all-or-none state. Thus, a lineup with an effective size of 4 is interpreted to mean that there were 4 plausible foils in the lineup. In the second place, mock witness choices are assumed to be random with respect to plausible foils. This assumption underlies much of the reasoning around lineup measures.

These assumptions are useful, and allow relatively easily executed tests of a line-up's worth. They also limit the way in which we evaluate lineups, and they lead, as we have seen, to interpretative difficulties.

I do not have carefully thought out alternatives to offer here, and can only point to the possible benefits of adopting two different assumptions. If plausibility is not viewed as an all-or-none state, but as continuously varying, then it makes better sense to evaluate foils for their degree of plausibility. The degree of plausibility is the extent to which foils fit the description of the perpetrator. In the case of the second assumption, we could relinquish the idea that we should view mock witness choice as random with respect to plausible foils, and think instead of foil identification proportions as estimates of foil plausibility. The proportion of mock witnesses identifying a foil could be thought of as an estimate of the degree of resemblance between the foil and the perpetrator, conditional on the resemblance the other foils bear to the perpetrator.

These revised assumptions have the benefit of unconfounding lineup size and lineup bias, but I am not sure how they could be implemented for practical benefit.

Diagnosticity and Informativeness

Despite the fact that the lineup has been used now for at least 100 years, its explicit purpose has rarely been carefully examined by courts. Some commentators assert that a lineup primarily provides protection against the suggestiveness inherent in other methods of identification such as a direct face-to-face confrontation between the witness and suspect (Devlin, 1976).[270] But what will the witness's positive identification show in such a situation? Will it show that the witness is reliable, and that the details of his/her testimony can therefore be taken seriously? Or will it provide direct evidence against the suspect, in the sense that it increases the (perceived) probability that the suspect is guilty?

Both of these aims are built into the structure of the lineup. The witness is asked to indicate the perpetrator[271] from a number of 'foils', who are known to be innocent of the deed in question. If the witness identifies one of the foils, his/her evidence regarding the identity of the perpetrator is considered less reliable than it would otherwise. The identification parade clearly serves on the one hand as a check on witness reliability. Yet, a positive identification is also taken as evidence against the suspect: indeed, the courts are at great pains to ensure that the identification at a parade constitutes

[270] It is worth noting here the recent research by Gonzalez, Ellsworth, and Pembroke (1993), which points out that no-one has ever provided empirical substantiation of this claim. Indeed, in several experiments conducted by Gonzalez et al. showups had a lower degree of response bias than lineups.

[271] Typically, the witness is also warned that the perpetrator may not be in the parade. In South Africa, for example, if a witness participating in a parade is not told this, then the parade will usually be considered irregular (Hoffman & Zefertt, 1989).

an independent piece of evidence.[272] This would not be the case if the identification parade served merely as a reliability check. Consistency of identification would serve as a measure of reliability in its own right. Instead, the courts prefer to consider identifications as independent evidence regarding the suspect's guilt.

How is the value of an identification obtained at a parade to be determined in terms of the second purpose attributed to it above? There are two distinct approaches to this question in recent psychological literature. Both approaches involve couching the problem in Bayesian terms, but they disagree sharply about the correct way to measure the probabilities involved. I will consider only the earlier of the approaches here, which derives from work by Wells and his associates (Luus & Wells, 1990; Wells & Lindsay, 1980; Wells & Turtle, 1986).[273]

The Diagnosticity Ratio

Wells and colleagues have devised two measures of the value of an identification. The first measure, 'diagnosticity', is defined as the amount of potential impact that an identification should have in revising one's opinion of guilt without regard to the prior estimate of guilt - how much more likely the data are to have occurred given the truth of one hypothesis (that the suspect is the criminal) relative to the other[274] (that the suspect is innocent):

$$\frac{P(ids|s = c)}{P(ids|s \neq c)}$$

where ids = identification; s = suspect; c = criminal.

Diagnosticity is thus a ratio of likelihoods: the ratio of the probability that the suspect is identified given that he is guilty, to the probability that the suspect is identified given that he is innocent. Note also that in order to obtain the diagnosticity measure, rates of identification are required from parades in which the criminal is present, and from parades in which the criminal is absent.

An example should make the meaning of the ratio clear. Table 6.8 presents the distribution of witness[275] choices in two sets of identification parades; in one of the parades in each set the suspect is guilty (parades l & n), in the other the suspect is innocent (parades m & o).

---

[272] See, for example, the South African case of *R* v *Kola*. Here the presiding juge, Schreiner, AJ. commented "[It is] unsatisfactory ... to rely upon ... evidence of identification given by a witness not well acquainted with the accused, if that witness has not been tested by means of a parade" (At 169).

[273] I referred briefly to the second of the approaches - Navon (1990a), Navon (1990b) - earlier in the chapter. I will not consider his position here.

[274] The hypotheses make up a *statistical partition* (i.e. they are mutually exclusive and exhaustive).

[275] Note that real witnesses, and not mock witnesses, are the witnesses of interest here.

	Lineup Member						
	1	2	3	4*	5	6	N/P
l	5	3	6	31	8	3	4
m	8	5	6	5	2	7	27
n	5	3	6	31	8	3	4
o	6	5	6	25	8	3	7

N/P = not present. * = Target.
Set {l, m}: diagnosticity = [31/60]/[5/60] = 6.20
Set {n, o}: diagnosticity = [31/60]/[25/60] = 1.24

Table 6.8 Diagnosticity in a series of hypothetical parades.

In the first set of parades, the suspect is 6.2 times more likely to be chosen when she is the criminal than when she is innocent. The identification parade is said to be diagnostic of the suspect's guilt. In the second set of parades, the suspect is only 1.24 times more likely to be chosen when guilty than when innocent, and the parade is thus not very diagnostic of the suspect's guilt.

The absolute size of the diagnosticity ratio is thus the measure of the line-up's value, taken independently of all other things that have a bearing on the case. It has been used extensively in the psychological literature in order to compare the relative success of structural alterations to lineup practice (Lindsay & Wells, 1985; Melara, De Wit-Rickards & O'Brien 1989). Valuable as the measure may be, little has been written about its statistical properties. This is a significant failing, which clearly hampers research in the area. In some cases, it appears to be misinterpreted,[276] and in others researchers have to devise ingenious research designs in order to test a difference in diagnosticity ratios for statistical significance.

There are several ways of conceptualizing the diagnosticity ratio that allow testing the ratio for statistical significance. I suggest one such model here.

In order to derive a diagnosticity ratio, a researcher conducts an experiment in which witnesses to a (staged) crime are randomly assigned an identification parade either containing the perpetrator or not containing the perpetrator. Diagnosticity is calculated as the ratio of the probability of identifying the suspect (also the perpetrator) in the perpetrator - present lineup to the probability of identifying the suspect in the perpetrator - absent lineup. Wells and colleagues apply a Bayesian analysis in order to evaluate the ratio of likelihoods, but I suggest an alternate conceptualization. The experiment can be represented as a cross tabulation, as in Table 6.9 below, where cell counts represent observed frequencies (n_{ij}):

[276] See Melara et al. (1989) for a case in point A diagnosticity ratio of less than one indicates that the lineup is of questionable value, since a positive identification is less likely than a mistaken identification, but Melara et al. appear to interpret a difference between two such ratios as evidence for the superiority of one of the lineups.

	Perpetrator Present	Perpetrator Absent	*Total:*
Identifies Suspect	n_{11}	n_{12}	n_{1+}
Identifies foil, or n/p	n_{21}	n_{22}	n_{2+}
Total:	n_{+1}	n_{+2}	n_{++}

Note that + notation is used to indicate summation across rows (r+) and columns (+c).

Table 6.9 Design used to derive empirical measures of diagnosticity.

The diagnosticity ratio is $\frac{n_{11}}{n_{+1}} \Big/ \frac{n_{12}}{n_{+2}}$, which is a ratio of (estimated) conditional probabilities. It is equivalent to an index widely used in biostatistics, called *relative risk*, since it expresses the probability that a guilty suspect is identified, relative to the risk that an innocent suspect is identified. When diagnosticity is 1.0, the events are equally likely; departure from 1.0, on the other hand, reflects the degree to which events are not equally likely.

I suggest one way of reasoning inferentially about the diagnosticity ratio. The method is applicable in the case of large samples, and employs corrections to increase approximation accuracy for small samples. Alternative methods are available, and these generally improve the approximation, but the labour involved in deriving the relevant confidence intervals is much greater. I refer interested readers to more comprehensive discussions by Agresti (1990), and Rothman (1986).

We assume that the columns of Table 6.9 represent independent binomial samples, with probability of success = $p_{1/i}$. This assumption allows us to determine an expression for the sampling error of the diagnosticity ratio, and to construct a confidence interval. I will not show the derivation of either here, since both results are reported elsewhere (Agresti, 1990).

The diagnosticity ratio is asymmetrical, and a log transformation to symmetry is therefore a useful preliminary step. Thus, where two lineups have diagnosticity = 6.24 and 0.16 respectively, the ratios are symmetrical under a log transformation (ln (6.24) = -ln (0.16) = 1.83). However, ln (d) will be undefined when either of n_{11}/n_{+1} or n_{12}/n_{+2} is zero. A correction for this is to add 0.5 to each of the frequencies in the calculation of the diagnosticity ratio:

$$d' = \frac{n_{11} + 0.5}{n_{+1} + 0.5} \Big/ \frac{n_{12} + 0.5}{n_{+2} + 0.5} \qquad (12)$$

A maximum likelihood estimator of the asymptotic standard error for d' is given by

$$S[\ln(d')] = \sqrt{\left(\frac{1}{n_{11}+0.5} - \frac{1}{n_{+1}+0.5} + \frac{1}{n_{21}+0.5} - \frac{1}{n_{+2}+0.5}\right)} \quad (13)$$

A 100*(1-α)% confidence interval for ln (d') is given by

$$\ln(d') \pm Z_{\alpha/2} \sqrt{\left(\frac{1}{n_{11}+0.5} - \frac{1}{n_{1+}+0.5} + \frac{1}{n_{21}+0.5} - \frac{1}{n_{2+}+0.5}\right)} \quad (14)$$

The endpoints of this interval can be exponentiated in order to transform the interval back into the metric of the diagnosticity ratio. The hypothesis that the ratio differs from 1 (i.e. d' ≠1, or, in log terms, ln (d') ≠ 0), can be tested at significance level α by determining whether the confidence interval contains 1.

Using (12), (13) and (14), the diagnosticity ratios presented in Table 6.3 have 95% confidence intervals of (3.98; 8.16) and (0.86; 1.77). Since the interval for the second ratio includes 1, we should perhaps treat the suspect-present lineup on which it is based as providing us with no more evidence about the suspect's guilt than the suspect-absent lineup.

Differences between independent diagnosticity ratios

Diagnosticity ratios are frequently used to compare the effect of independent manipulation of lineup characteristics. Thus, Lindsay, Lea & Fulford (1991) compared diagnosticity ratios of lineups in which the presentation of lineup members was sequential or simultaneous, or a combination of both. The difference between diagnosticity ratios will of course be subject to random sampling variation, just as individual diagnosticity ratios are. Comparisons would be aided by placing some probabilistic confidence in the size of the difference.

In this section I suggest a test for the difference of two diagnosticity ratios, or the homogeneity of more than two diagnosticity ratios. The approach I take is based on that outlined by Rothman (1986, pp 220-233) apropos of assessing the homogeneity of relative risk effects.

Imagine that there are k independent diagnosticity ratios. Each diagnosticity ratio (d) is transformed to ln (d).[277] Then the approximate variance of ln (d) is given as

$$ar[\ln(d_i)] = \frac{n_{21}}{n_{11}n_{+1}} + \frac{n_{22}}{n_{12}n_{+2}} \quad (15)$$

[277] Note that the addition of 0.5, used as a correction in (12), is not used here. The method is therefore only applicable to sets containing non-zero diagnosticity ratios. A simple correction should be feasible, but is not investigated in this chapter.

where the notation is the same as that used in Table 6.10, and i is the i^{th} diagnosticity ratio.

In order to obtain a pooled estimate of the diagnosticity ratios, we use a weight for each ratio that is equal to the inverse of its variance:

$$w_i = \frac{n_{11}n_{12}n_{+1}n_{+2}}{n_{11}n_{22}n_{+1} + n_{12}n_{21}n_{+2}} \tag{16}$$

Then the pooled estimator is

$$\bar{d} = \exp\left[\frac{\sum_{i=1}^{k} w_i \ln(d_i)}{\sum_{i=1}^{k} w_i}\right] \tag{17}$$

The homogeneity of the set of diagnosticity ratios can be tested by computing the summed squared distance of each ratio around the pooled estimate, where individual ratios are weighted by their variances. The result is a χ^2 deviate, with k-1 degrees of freedom, i.e.

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} \frac{[\ln(d_i) - \ln(\bar{d})]^2}{\text{var}[\ln(d_i)]} \tag{18}$$

Better approximations are available, but at the cost of considerable computational effort. Rothman (1986) outlines several alternatives, based largely on maximum likelihood methods.

By way of example, Table 6.10 reports three diagnosticity ratios obtained by Lindsay et al. (1991), and the results of a test for homogeneity of the ratios. (The test for homogeneity is provided as an example only - the diagnosticity ratios determined by Lindsay et al. were not independent).

	Lineup Method		
	Simultaneous	Sequential	Simultaneous & Sequential
Diagnosticity (d)	2.84	9.34	2.58
ln (d)	1.04	2.23	0.95
Var ln(d)	0.16	0.18	0.04

$\bar{d} = 3.19$; $\chi^2 = 7.55$, $p < 0.025$

Data from Lindsay et al. (1991).

Table 6.10 A test of the homogeneity of three diagnosticity ratios

The analysis suggests that the differences between the three diagnosticity ratios are not simply a function of random sampling variation. Standardized residuals could be examined in order to identify particularly salient deviations from the pooled estimate: in Table 6.10 it is clear from the individual diagnosticity ratios that the sequential parade has the largest associated ratio.

In the absence of an inferential basis the comparison of diagnosticity ratios is a questionable business. Differences in absolute magnitude may arise from sources other than real differences in lineup technique: framing the comparison in probability terms goes some way toward bolstering comparisons.

Information Gain

A second measure rooted in Bayesian thinking, devised by Wells and colleagues, and closely related to diagnosticity, is known as 'information gain'. A positive (or negative) identification presumably leads to some alteration of the probability that the suspect is the criminal from the point of view of the investigating officer, and the amount of this change is the nett informational value of the identification. Information gain is the difference between the prior probability that the suspect is the criminal, and the posterior probability of the same (dependent only on the intervening identification or non-identification of the suspect). It is stated formally as

$$\text{Information gain} = |p(s=c) - p(s=c|ids)|$$

$$\text{or,} \quad \text{Information gain} = |p(s\neq c) - p(s\neq c|nids)|$$

where s = suspect, c = criminal, | = sign for conditional occurrence of an event, ids is the event that an identification is made, and nids is the event that an identification is not made;

$$p(s = c|ids) = \frac{p(ids|s = c)p(s = c)}{p(ids|s = c)p(s = c) + p(ids|s \neq c)p(s \neq c)},$$

and s, c, ids are defined as before.

The absolute value of the difference between the probabilities is taken, since it is the size of the gain that is important, and not the direction. The value of the identification depends on the size of the prior probability (although this is only clear from a close examination of the Bayesian ratio): a positive identification might lead to a big increase in the likelihood of the suspect's guilt if it is low to begin with, but will generally have a smaller impact the higher the prior likelihood of guilt.

The measure of information gain has proved most useful as a research tool: thus Wells & Lindsay (1980) used it to show that non-identifications must be diagnostic of innocence, just as identifications are diagnostic of guilt, and Wells & Turtle (1986) used it to show that multiple suspect lineups will, in general, be inferior to single suspect lineups.

Although the measure of information gain is rooted in a Bayesian analysis, I suggest that it is possible to apply non-Bayesian inferential reasoning to it with an appropriate re-conceptualization. One re-conceptualization is to set the prior probability to expectation under the assumption of equiprobability, and then to regard any difference of the observed probability of identification from expectation as information gain. This is equivalent to assuming that the suspect is innocent before the lineup, and that the extent to which he is chosen at a rate exceeding chance expectation indicates probability of guilt. This conceptualization has the advantage of permitting probabilistic statements regarding the size of information gain.

In practical terms, a useful way of implementing this conceptualization may be to calculate a confidence interval (for some choice of α) around the rate at which the suspect is chosen, using (3) or (4) above. The interval can then be corrected by subtracting the expected rate of identification from each of the extrema. The corrected interval expresses information gain as a $100*(1-\alpha)$ percent confidence interval. Intervals that include 0 could be considered to support a hypothesis of zero information gain. It is possible that the interval will be entirely negative: the meaning of this will depend on the nature of the lineup task. If the lineup is constructed to resemble a description of the perpetrator, then it conveys information suggesting a lack of resemblance between the suspect and the perpetrator. If the lineup is constructed to resemble the suspect, then an entirely negative interval will be more difficult to interpret.

As an example of this kind of calculation, consider the sequential lineup condition from the study by Lindsay et al. (1991). In the perpetrator - present lineup, 28 of 60 mock witnesses identified the suspect (perpetrator), from a lineup with 8 members. A 95% confidence interval around this proportion is (0.35; 0.61). If we subtract the expected proportion from the endpoints, the new interval is (0.22; 0.49), which we can think of as the 95% confidence interval estimate of information gain. This interval suggests that the lineup yields positive identifying information regarding the guilt of the suspect, *ceteris paribus*.

Discussion and conclusion

I have made several suggestions in this chapter aimed at sharpening the usage of existing measures of lineup fairness and informativeness. There is little formal consideration in the psychological literature of the application of inferential reasoning to lineup tasks, and yet most measures of fairness derive their meaning from a probabilistic assumption. This failure is significant.

Measures of lineup fairness are generally based on the mock witness task, devised by Doob and Kirshenbaum (1973). This task requires that subjects blind to the identity of the suspect attempt to guess the suspect from an array of lineup members. Inferences are made regarding the fairness of the

lineup depending on the departure of the suspect identification probability from expectation, which is determined under an assumption of equiprobablity: that is, mock witness identifications are assumed to be random across foils. Since mock witness identifications are assumed to be random, the number of identifications of the suspect will show the effects of random sampling variation. This is as true of the mock witness task as it is of a coin tossing experiment. Just as we would be loathe to interpret the outcome of a coin tossing experiment without some explicit probability theory, so we should hesitate before interpreting the outcome of mock witness tasks. We should also hesitate before interpreting measures of fairness derived from the task, which include the widely used 'functional' and 'effective' sizes.

I have suggested ways here of thinking inferentially about these lineup measures. In the first place the mock witness task is conceptualized in terms of a simple binomial probability model, which allows the calculation of exact probabilities associated with identifications of the suspect. This conceptualization allows us to postulate inferential methods for the interpretation of measures of functional and effective size. Without these methods, there are problems for both research and application. Estimates of functional and effective size have been used to evaluate the fairness of lineups used in basic research, but it is not clear that fairness can be evaluated in this way without an appropriate theory of its statistical variation. This problem is emphasized in applied legal settings, where the measures in question have been used by expert witnesses to support fairness evaluations of police lineups (see Buckhout et al., 1988, for an example). If an expert witness claims, on the basis of a mock witness evaluation of a police lineup, that the effective size of the lineup is 6, this may be misleading, particularly if the sample of mock witnesses is relatively small, and the lineup size is relatively large. I have suggested ways in this chapter in which estimates of functional and effective size are re-phrased as confidence intervals. This would not, in my opinion, confuse jury members or legal personnel, since quantification and reporting of measurement error are commonplace in the press, particularly in relation to opinion polls. It would have the benefit, though, of making evaluations of lineup fairness more rigorous, especially since these evaluations would be tied more directly to the probability models implicitly underpinning them.

There may be better ways of achieving what I have attempted here. In particular, it may be possible to develop a unified approach to the several measures of fairness and informativeness; my goals in this chapter were simply to suggest useful inferential methods, which are neither computationally intensive nor dependent on advanced statistical knowledge.

What would certainly be of considerable benefit, though, is a close scrutiny of some of the assumptions underlying the notions of effective and functional size. There seem to be certain conceptual problems attached to the claim that a lineup consists of k plausible foils, and I argued earlier that these lead to the confounding of lineup size and lineup bias. Plausibility is treated in most

research as an all-or-none state - this is what gives meaning to claims like 'this lineup has 4 plausible foils' - but it is more likely that plausibility is a matter of degree than a discrete state. Navon's remarks on the 'proper treatment of diagnosticity' (1990a, 1990b) may prove a fruitful line of reasoning in this regard, since his suggested conditional probability model explicitly regards the value of a suspect identification as conditional on the degree of resemblance between suspect and perpetrator.

Most of the measures of lineup fairness and informativeness were developed under the assumption of a simultaneous lineup structure: that is, witnesses view an array of people, presented simultaneously, and attempt to choose the perpetrator, if they feel he/she is present. Lindsay, Wells and colleagues have shown very convincingly that if the structure of the task is changed so that the array is presented sequentially, then witness performance is dramatically better (Lindsay & Wells, 1985; Lindsay et al., 1990).[278] There is some evidence that U.S. police are starting to use the sequential structure more regularly: perhaps even as many as 15% of lineups in some areas are now sequential (Malpass, personal communication).

It is not clear that the measures developed under the assumption of simultaneous lineup structure will be useful for the evaluation of fairness in sequential lineups, and it is therefore also not clear that the inferential considerations developed in this chapter will be applicable to sequential parades without substantial modifications.

The mock witness task could still be used to gauge fairness of sequential lineups if the procedure used to implement the test meets certain conditions. Consider the case of a lineup sequence in which all the foils are perfectly fair. We would expect mock witnesses to choose early in the sequence if they are using the criterion of resemblance fairly (they are required to identify a lineup member if that person matches the description). It follows that if the sequence is unaltered for other mock witnesses - i.e. if the same foils are always presented first, second, and so on - then the distribution of mock witness choices will be highly skewed, in favour of the early foils. To ensure a fair distribution of identifications across the lineup, one would need to completely counterbalance the presentation of foils. Unfortunately this would be impracticable for lineup sequences of any significant length (a 10 member lineup would require 3.63×10^6 sequences). One would instead need to use a subset of possible sequences. Such a subset could consist of a random selection of possible sequences, but there would need to be a fair number of these in order to ensure that particular foils do not appear early in the sequence more frequently than other foils. By the same reasoning, a substantial number of mock witnesses would be required, since each sequence would need to be well tested.

[278] There are several other structural aspects to the sequential lineup, some of which appear to have become standard in reports on their use. One of these is that the witness does not know the number of lineup members.

The point behind these procedures is to ensure that a completely fair sequential lineup results in a uniform distribution of identifications, since departures from uniformity could then be used as a gauge of the fairness of the lineup, and in particular, as reflections of the plausibility of individual foils. The measures of functional and effective size depend on the assumption that uniformity of identifications is the theoretically 'fair' state against which departures should be judged, and if a sequential mock witness task can satisfy this assumption, then the measures should be equally useful for this kind of lineup. As we have seen, however, there are clearly several practical difficulties in constructing a version of the mock witness task for sequential lineups. But, more importantly, such a mock witness task would lack the clear interpretability of the task used for the simultaneous lineup. In the simultaneous lineup, expected performance is easy to determine, since witness choice is assumed to be random across the k members. It is not clear what expected performance is in the sequential lineup, since witnesses do not know the number of lineup members, and choice is unlikely to be random across k members. This is a significant failure, since the elegance and power of the mock witness task depends on the simple probability model underlying it. It is a simple matter in simultaneous lineups to determine whether a suspect is chosen more frequently than chance, but in sequential lineups the quantitative nature of random choice will be very difficult to estimate.

A possible way of approaching the problem is to use simultaneous lineups to test the fairness of sequential lineups (the lineups would, of course, have the same members!) It is perhaps reasonable to assume that fair simultaneous lineups are also fair sequential lineups,[279] and that foils who are plausible in a simultaneous lineup will also be plausible in a sequential lineup. This implication is probably only true in one direction, since there is some evidence that sequential lineups are relatively robust with respect to problems of low functional and effective size (Lindsay et al., 1991). Fair sequential lineups are therefore not necessarily fair simultaneous lineups.

[279] I refer only to the structure of the lineup, of course. Procedural problems in a sequential lineup could well render a simultaneous lineup unfair.

Chapter 7 ~ Preliminary Empirical Work

In this chapter I commence the empirical exploration of a measure of facial similarity derived from the principal component analysis (PCA) of a set of facial images. The measure was discussed at some length in Chapter 4, so there is no need to re-trace the theoretical and mathematical justification, except where the empirical results suggest that this is necessary. The chief goal of the empirical exploration here is to gather evidence on the correspondence between the PCA similarity measure and human judgements of facial similarity. Two studies are reported in this chapter: in the first, a small set of faces is analysed, and performance on a similarity ranking task is compared to performance predicted by the PCA measure. In the second study, a larger set of stimulus faces is used, and four validity measures are compared against performance predicted on the basis of the PCA measure. The PCA measure shows considerable promise, but there are some mathematical and theoretical questions that need addressing. In addition, the similarity judgement tasks are problematic in several respects, and need re-thinking.

Study 1

Study 1 served as a pilot investigation. Two small sets of faces were selected from a larger collection, and were composed as an array, or 'simultaneous photospread lineup'. The analogy of a photospread was explicit in the composition. One of the chief potential applications of the measure of facial similarity is lineup evaluation: judgements of facial similarity made by witnesses in this situation are inherently comparative,[280] and a similarity measure derived from the principal component representation of the photospread is intuitively satisfying.

Method

Stimuli

Two sets of six facial images from the set collected in Study 2 were used in this study. Details regarding the method of image collection and standardization are provided in the description of Study 2. Six white male faces constituted set 1a, and six black female faces constituted set 1b. The faces were not chosen on the basis of particular criteria, but simply to make up arrays that were relatively homogeneous with respect to gross external attributes. Each set was printed on plain paper at 360 dpi

[280] This is the case for simultaneous lineups, but may not be the case for sequential lineups - see Chapters 4 and 6 for a discussion of the differences between these lineups.

(dots per inch), using an inkjet printer, for presentation to subjects. Appendix B shows the sets of faces. Three sequences were created for each set, for purposes of counterbalancing.

Analysis of stimuli

The six images in each stimulus set were standardized in the manner described for the stimulus set of Study 2, except that photographs were scanned at size = 550 x 600 pixels. These image sets were then analysed in the same manner described for Study 2, except that face images were 330 000 element vectors. Image set 1a (white male faces) and the corresponding eigenfaces generated by PCA are represented in Chapter 5 as Figures 5.7 and 5.8. Both image sets are shown as Appendix B. Table 7.1 presents a summary report of the principal component analysis.

Image set 1a (white male)				Image set 1b (black female)			
PC no.	Eigenvalue	% Variance	Cum. %	PC no.	Eigenvalue	% Variance	Cum. %
1	4.36	72.6	72.6	1	3.78	62.9	62.9
2	0.55	9.2	81.8	2	0.87	14.4	77.4
3	0.36	6.1	87.8	3	0.82	13.6	91.0
4	0.33	5.5	93.3	4	0.26	4.3	95.3
5	0.22	3.7	97.0	5	0.17	2.8	98.1
6	0.18	3.0	100.0	6	0.12	1.9	100.0

PC = principal component. % Variance = percentage variance explained by component. Cum. % = cumulative percentage of variance explained by all preceding components, in addition to the present component.

Table 7.1 Principal component analysis of images in sets 1a and 1b

It is clear from Table 7.1 that both image sets produce first principal components which resolve a considerable amount of variance. Several popular statistical 'rules of thumb' would suggest that the first component is sufficient: all other components have eigenvalues smaller than 1 (Kaiser, 1958), and a scree test (Cattell, 1966) in both cases shows an 'elbow' after the first component. However, the image sets are small, and the components may be quite unstable. It is also possible that dimensional reduction of the component solution will lose important information about the physical similarity of images. In the analyses reported below, the full solution is used to generate similarity scores. (The issue of dimensional reduction is explored at greater length in Study 2).

Subjects

Subjects were 35 school pupils attending an evening discussion on career possibilities at a local high school. Their ages ranged from 16 to 18.

Procedure

Subjects were asked at the end of a career evening to participate 'in a study on face perception and recognition'. Each subject was given two sets of faces, corresponding to those described above under **Stimuli**. These sets were combined in an experimental booklet, in alternating order across subjects (i.e. some subject received booklets which presented the male array first, and others received booklets which presented the female array first). Facial images were arranged in varying sequences within arrays, as described above, to effect counterbalancing. Subjects were (verbally) asked to rank the faces in order of similarity to a 'target face' by placing the number '1' beneath the most similar face, '2' beneath the next most similar face, and so on. The target was indicated by enclosure in a rectangle, and by the absence of prior numbering. Subjects were given fifteen minutes to complete the task, after which booklets were collected.

Results

Rankings provided by subjects were analysed chiefly for correspondence to the similarity metric derived from the PCA, but several other issues were also explored.

Orders and sequences

Image sets 1a and 1b were combined in several different *orders* (i.e. the male and female arrays varied with respect to whether they were presented first or second), and in different *sequences* (i.e. faces within the array varied with respect to the position in the array), for counterbalancing purposes. In order to test the effect of the different orders and sequences on similarity rankings, a multivariate analysis of variance (MANOVA) was conducted, taking orders and sequences as between subjects effects, and rankings of the separate faces as a within subjects effect. Table 7.2 shows that the between subjects effects and all interactions involving these effects, were not statistically significant. All further results in Study 1 will therefore be reported without respect to order or sequence of presentation. The within subjects effect ('RANKING' in the table) is significant, though, which indicates that faces were ranked differentially in terms of similarity to the target face.

Set 1a

Tests of Between-Subjects Effects.

Source of Variation	SS	DF	MS	F	p <
WITHIN + RESIDUAL	.31	30	.01		
ORDER	.01	1	.01	.77	.387
SEQ.	.02	2	.01	.86	.434
ORDER BY SEQ.	.02	2	.01	.86	.434

Tests involving Within-Subject Effect.

Source of Variation	SS	DF	MS	F	p <
WITHIN + RESIDUAL	179.67	120	1.50		
RANKING	120.21	4	30.05	20.07	.000
ORDER BY RANKING	9.49	4	2.37	1.58	.183
SEQ. BY RANKING	18.42	8	2.30	1.54	.151
ORDER BY SEQ. BY RANKING	12.25	8	1.53	1.02	.423

Set 1b

Tests of Between-Subjects Effects.

Source of Variation	SS	DF	MS	F	p <
WITHIN + RESIDUAL	2.74	28	.10		
ORDER	.06	1	.06	.62	.438
SEQ.	.12	2	.06	.63	.538
ORDER BY SEQ.	.12	2	.06	.63	.538

Tests involving Within-Subject Effect.

Source of Variation	SS	DF	MS	F	p <
WITHIN + RESIDUAL	256.65	112	2.29		
RANKING	46.24	4	11.56	5.05	.001
ORDER BY RANKING	8.08	4	2.02	.88	.478
SEQ. BY RANKING	19.38	8	2.42	1.06	.398
ORDER BY SEQ. BY RANKING	17.81	8	2.23	.97	.462

ORDER = order of presentation of sets 1a and 1b; SEQ. = arrangement of faces within the sets

Table 7.2 Analysis of variance table for effects of orders, sequences, and similarity rankings of faces

Consistency of rankings

Although the analysis immediately above shows that subjects, on average, reliably differentiated faces in terms of similarity to the target, the rankings showed considerable inter-subject variation. The degree of correspondence was investigated by calculating Kendall coefficients of concordance, and testing these for statistical significance. The size of these coefficients indicates the degree of concordance, and the test of statistical significance estimates the probability that this degree of concordance is a product of mere random sampling variation. Coefficients were calculated on rankings for each of the image sets, and these were statistically significant, but low ($w = 0.40$, $\chi^2 = 57$, $df=4$, $p < 0.0001$, for set 1a; $w = 0.13$, $\chi^2 = 17.58$, $df=4$, $p < 0.001$, for set 1b). Table 7.3 attempts to demonstrate the degree of overall concordance, by cross-tabulating faces and ordinal positions in the ranking sequence.

		Set 1a					Set 1b				
		A1	A4	A2	A5	A3	B1	B5	B3	B4	B2
Ordinal position	1	25	2	3	5	2	12	12	5	2	3
	2	7	9	12	5	7	8	7	8	6	5
	3	2	12	3	11	13	10	3	10	8	3
	4	1	8	8	5	13	3	5	5	10	10
	5	0	4	9	9	0	1	7	6	8	12
Mean Rank		1.4	3.1	3.2	3.3	4.2	2.2	2.7	2.9	3.5	3.6
Std. dev.		0.7	1.1	1.4	1.4	0.9	1.1	1.6	1.3	1.2	1.5

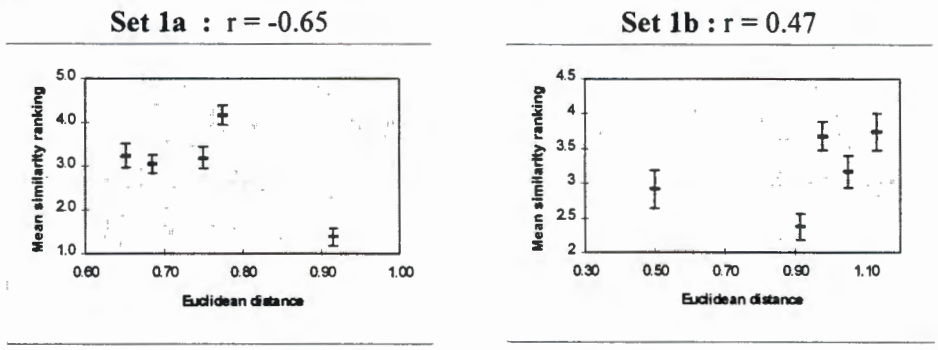
Note: cell entries represent frequencies of rankings of faces in the image sets. A1 denotes the 1st face in set 1a, A2 the second face, and so on.

Table 7.3 Similarity rankings of faces in image sets 1a and 1b

This method of presenting subject rankings has the following implications: perfect concordance would result in diagonal matrices (i.e. entries only in the main diagonal), and states of lesser concordance would be shown by the departure of matrices from this condition. Although frequencies in diagonal and off-diagonal cells are discernibly larger than in other cells, this is more clearly the case for image set 1a, and the pattern in both cases is far from near-diagonal. However, the mean rankings show that subjects do not clearly differentiate several faces in terms of their similarity to the target (e.g. faces A4, A2 and A5), and it may be that these faces are not sufficiently dissimilar (from each other) to elicit concordant rankings from subjects.

Similarity correspondences

Similarity ratings obtained from subjects were correlated against similarity scores derived from the PCA described in **Stimuli**, above. The correspondence is difficult to interpret, either from the scatterplots, or the correlation coefficients, which are shown in Figure 7.1.



Similarity rankings for sets 1a and 1b were provided by 35 and 34 subjects, respectively. Error bars are extended one standard error above and below the mean.

Figure 7.1 Relations between ranked similarity of faces in sets 1a and 1b, and distances derived from PCA.

Although it appears from the scatterplots that there is only a small number of data points across which the correlation is calculated, and that this might make the magnitude of the relation difficult to interpret, it should be remembered that each of the data points represents a mean ranking, calculated over more than 30 subjects. However, these means are based on scores which exhibit considerable variability, and the estimates they give may be quite inaccurate. The difference in the direction of the relation across the image sets is surprising, and may be a consequence of the instability. In addition, it is apparent that the faces in the image sets do not show much between-face variation: subjects, on average, rate the five faces as bearing much the same similarity to the target face, and the Euclidean distances derived from the PCA also show a restriction in range of variation. This is a flaw in the design of the rating task: faces were not chosen to represent different degrees of similarity to the target, and it would probably have been wiser to select the faces on the basis of Euclidean distance from the target.

Similarity measures derived from partial faces

In some earlier work which used PCA to generate face subspaces, hair was excluded from the analysis by digitally masking faces (Sirovich & Kirby, 1987; Kirby & Sirovich, 1990). This is often the case in research on face perception processes, presumably on the ground that hair is an adaptable property of faces, and judgements which use hair properties do not reflect judgements about the faces per se (see Bruce & Healey, 1991 for example). In order to test the effect masking has on similarity scores derived from PCA, the six faces of set 1a were digitally 'masked' by cropping them just above the eyebrows, and next to the eyes (see Figure 7.2), and these 'masked' images were subjected to PCA. Figure 7.2, which reports the correspondence between similarity scores determined from full faces and those determined from masked faces, shows that masking had little effect on the similarity scores.

Sample masked face

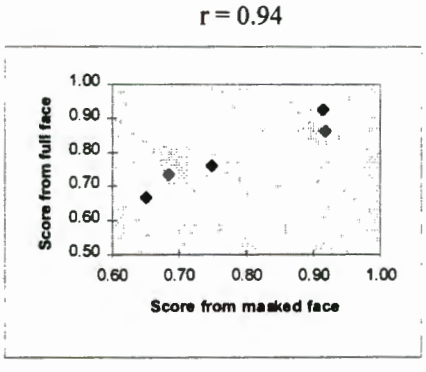


Figure 7.2 Correspondence between similarity scores derived from PCA of full faces, and similarity scores derived from PCA of masked faces.

Similarity measures derived from larger image subspaces

Similarity measures reported thus far were derived from analyses of arrays of six faces. Although this is a restriction based on practical considerations - a prospective application of the measure is to lineups, which usually have small numbers of members - it is not preferable from a theoretical point of view. The similarity measure is calculated on the basis of dimensions identified by the PC analysis; these dimensions will differ when PCA is conducted on larger sets of faces, and lower dimensions will be much more stable. In order to empirically examine the effect of using similarity scores derived from larger subspaces of faces, the faces in set 1a were included in a larger set of 62 (the set reported in Study 2 below), and similarity scores derived from the PC solution for this set. Figure 7.3 shows that the similarity scores calculated from the PCA of six faces did not differ markedly from those calculated from the PCA of the larger set.

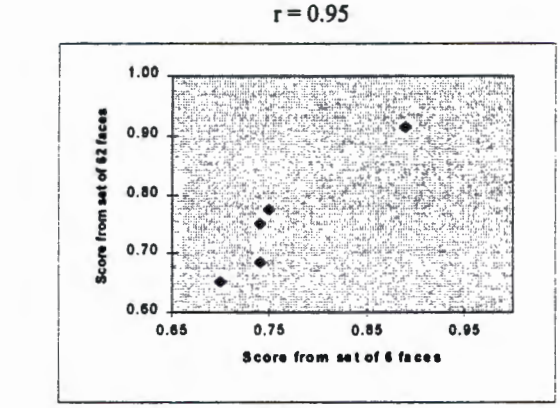


Figure 7.3 Correspondence between similarity scores derived from PCA of set 1a = six faces, and similarity scores for the same six faces derived from PCA of a set of 62 faces.

Discussion

Study 1 does not provide any clear evidence to support the notion that the proposed measure of similarity corresponds to judgements of similarity made by human subjects. It also does not provide any clear evidence against this notion, though. Several weaknesses inherent in the judgement task used in Study 1 render it incapable of providing convincing evidence of either kind. Only five faces were used in the task, and although this is not a problem in itself (since more than 30 responses are pooled to provide a similarity estimate for each face), when coupled with high inter-subject variability, the task may provide seriously inaccurate estimates of perceived similarity. Three amendments to the design may alleviate the problem: (1) constructing the arrays so that faces vary systematically and substantially on the PCA measure of similarity; (2) using more subjects, so that

obtained mean rankings are more reliable; and (3) finding alternate similarity rating tasks. These amendments are implemented in Study 2, below.

The substantial variation in perceived similarity is an interesting finding in its own right. The phenomenon is unreported in previous research, apart from a casual observation made by Lindsay (1994), which is surprising. It raises a considerable problem: if perceived similarity shows little correspondence across subjects, then the task of constructing a lineup of people who meet a criterion level of facial similarity may be impossible. Possible explanations should be considered, though. Firstly, it may be that people will agree only when faces show substantial similarity or dissimilarity, but not when they show levels of similarity between the extremes. This is explicable if judgements of similarity are made on multiple bases - that is, if they are multi-dimensional - and if these differ across subjects. Highly similar faces will resemble each other on many dimensions, and judgements based on a subset of these dimensions are more likely to be congruent. Highly dissimilar faces will resemble each other on very few dimensions, and will be likely to attract ratings of dissimilarity. Other faces are problematic: since they will resemble each other on some dimensions, but not on others, there is considerable room for disagreement when subjects use only a subset of available dimensions. This hypothesis can be tested by examining the variation shown in subject ratings of similarity: variability should be lowest for faces rated most- and least- similar, and higher for other faces. Table 7.3 bears this contention out (more clearly in the case of set 1a), but the differences in variation are not compelling. Another possibility is that subjects are themselves simply inconsistent in rating similarity. A useful test of this might be the investigation of the consistency of subject ratings on different occasions.

Study 2

In Study 2, further investigations were conducted on the validity of the PCA measure of facial similarity. The correspondence between rated similarity and the PCA measure was again investigated: modifications were effected to the task reported in Study 1, and a new rating task was constructed. The modifications to the rating task included i) the careful selection of faces, aimed at ensuring adequate variation of PCA-based similarity scores between faces, and ii) the inclusion of a greater number of target faces in arrays. The new rating task required subjects to match pairs of faces in terms of similarity, rather than ranking an array of faces in relation to a target.

Three additional tests of the validity of the PCA similarity measure were explored in Study 2. In the first place, the ability of the PCA similarity measure to predict the race, sex, and age of faces was tested. If the measure conveys information about the similarity of faces, it should be able to discriminate clearly dissimilar groups. This predictive ability was investigated with discriminant function analysis.

In the second place, the relation between a PC-based distance measure and rated facial distinctiveness and typicality was explored. In the theoretical justification of the measure in Chapter 5, I proposed that the measure should capture distinctiveness information, and this proposition was accordingly investigated in Study 2.

In the third place, the face validity of PCA similarity scores was investigated. This was achieved by subjecting similarity scores to a cluster analysis: faces constituting the ensuing clusters were examined for physical similarities. The rationale behind this procedure was that similarity scores will possess some 'face validity' if clusters identified by a cluster analysis appear to be differentiated on the basis of facial 'features'.

Method

Stimuli

Seventy two students and staff at the University of Cape Town (UCT) were photographed against a uniform, dark background. These photographic subjects were recruited by placing advertisements on noticeboards on the UCT campus, by advertisement in lectures, and by direct recruitment as they passed the venue in which photographs were taken. Subject characteristics are reported in Table 7.4

Age		**Sex**		**Race**[281]	
Category	N	Category	N	Category	N
18 - 29	45	Male	21	Black	8
30 - 39	6	Female	41	Coloured	13
40 - 49	7			White	41
> 50	4				

Table 7.4 Subject characteristics of the 62 faces submitted to PCA analysis in Study 2.

Ambient lighting was standardized,[282] a stroboscopic flash unit was used to provide a direct source of light, and was coupled to an automatic 35mm Leica SLR camera, which was used to capture photographs. Subjects were asked to adopt a neutral expression, and to look straight ahead at the camera. Photographs were developed by a commercial photographic service, and then digitally scanned at 300 dpi on a Hewlett Packard IIx flatbed grey-image scanner. The images were edited

[281] Race groups reported here are based on those defined in the (now defunct) South African Population Registration Act. They should not be taken to indicate distinct genetic or physiognomic populations, although the groups do differ considerably in physical appearance.

[282] All electric lights in the venue, which was a seminar room, were turned on. These consisted of two neon lights, and six incandescent lights. The effect of variations in background and lighting on component based facial representations was investigated by Sirovich & Kirby (1990), who concluded that the effect is small.

digitally to remove jewellery and other extraneous costume. Of an original set of seventy two photographs, ten were rejected: five were rejected due to poor quality,[283] and five were rejected because subjects had facial hair, or spectacles which were not removed at the time of taking photographs.

Analysis of stimuli

The sixty two images in the stimulus set were standardized with respect to position of the left and right pupils (i.e. images were cropped, enlarged, or reduced so that the pupils occupied the same co-ordinate positions in a common pixel space). Image size was equated by cropping to a uniform size of 154 x 205 pixels. Each face was therefore standardized to a vector of size = 31570 pixels. (The set of images is reproduced as Appendix C). The image set was then submitted to PCA, using SPSS™ proprietary software under an MS Windows™ platform. Face images were submitted as variables, each constituted by 31570 'observations'. Principal components and their coefficients were derived from this analysis, and these were used to generate a matrix of Euclidean distances between faces in the image set (61 of the principal components were used in one set of analyses, and 5 in another set). In addition, the Euclidean distance of each face from the 'centre'[284] of the component space was calculated, in order to investigate the relation between rated distinctiveness, typicality, and distance of faces from the centre of the component space. The principal component analysis is reported in summary form in Table 7.5

PC no.	Eigen-value	% Var	Cum %	PC no.	Eigen-value	% Var	Cum %	PC no.	Eigen-value	% Var	Cum %
1	32.08	51.7	51.7	22	0.29	0.5	90.3	43	0.13	0.2	97
2	8.85	14.3	66	23	0.27	0.4	90.8	44	0.13	0.2	97.2
3	2.99	4.8	70.8	24	0.26	0.4	91.2	45	0.13	0.2	97.4
4	2.25	3.6	74.5	25	0.26	0.4	91.6	46	0.12	0.2	97.6
5	1.10	1.8	76.2	26	0.26	0.4	92	47	0.12	0.2	97.8
6	1.06	1.7	77.9	27	0.25	0.4	92.4	48	0.12	0.2	98
7	0.86	1.4	79.3	28	0.24	0.4	92.8	49	0.12	0.2	98.2
8	0.82	1.3	80.6	29	0.22	0.4	93.2	50	0.11	0.2	98.4
9	0.68	1.1	81.7	30	0.21	0.3	93.5	51	0.10	0.2	98.5
10	0.58	0.9	82.7	31	0.20	0.3	93.8	52	0.10	0.2	98.7
11	0.53	0.9	83.5	32	0.20	0.3	94.2	53	0.10	0.2	98.8
12	0.50	0.8	84.3	33	0.19	0.3	94.5	54	0.09	0.2	99
13	0.47	0.8	85.1	34	0.18	0.3	94.8	55	0.09	0.1	99.1
14	0.45	0.7	85.8	35	0.18	0.3	95	56	0.09	0.1	99.3

[283] The photographic negatives for these images were (erroneously) exposed to contaminating light by the photographic processing service.

[284] The 'distance from the origin' is an intuitive way of thinking about facial distinctiveness, but in practice this measure cannot be used. Principal component analysis 'normalizes' vectors, with the consequence that the Euclidean distance of each face from the origin is 1. Another way of conceptualising distinctiveness is as the distance of each face from the multivariate mean of component coefficients: typical faces will be close to the 'average tendency', and distinctive faces far away.

15	0.41	0.7	86.5	36	0.17	0.3	95.3	57	0.09	0.1	99.4
16	0.39	0.6	87.1	37	0.17	0.3	95.6	58	0.08	0.1	99.5
17	0.38	0.6	87.7	38	0.16	0.3	95.8	59	0.08	0.1	99.7
18	0.36	0.6	88.3	39	0.16	0.3	96.1	60	0.07	0.1	99.8
19	0.33	0.5	88.8	40	0.15	0.2	96.3	61	0.07	0.1	99.9
20	0.32	0.5	89.4	41	0.14	0.2	96.6				
21	0.30	0.5	89.9	42	0.14	0.2	96.8				

See Table 7.1 for key.

Table 7.5 Principal component analysis of images in Study 2.

The sizes of the principal components (five > 1.0) suggest that a five dimensional solution would be adequate here (Kaiser, 1958), but a stricter criterion (Joliffe, 1972) suggests an eight dimensional solution. A scree plot, reproduced below as Figure 7.4 appears to provide support for a five dimensional solution.

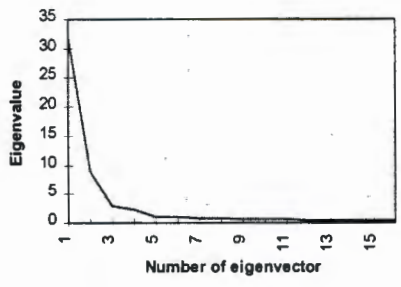


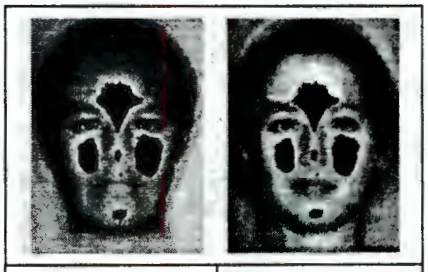
Figure 7.4 'Scree' plot for principal components of images in Study 2.

In similar, previous research using principal components analysis on face images, O'Toole et al. (1994) have claimed that early components 'capture' information about the race and sex of images. Figure 7.5 reproduces the first ten 'eigenfaces' derived from the present analysis, and subjective scrutiny of the images appears to bear this contention out.

Subscripts index the ordinal sequence of the eigenfaces; thus 'E1.GIF' is the first eigenface, and 'E10.GIF' the tenth eigenface. Eigenfaces 2 to 10 have been brightened for presentation purposes. The black 'patches' on the faces (particularly E1) may be an artefact of stroboscopic illumination.

Figure 7.5 The first ten 'eigenfaces' derived from the PCA of face images in Study 2.

The first eigenface appears to represent the 'central tendency' (or, perhaps 'faceness') of the images, whereas eigenfaces 2 and 3 appear to represent the race and sex of images. It should be remembered, though, that summed weightings of these eigenfaces re-create the images in the original set of faces, and it is probable that combinations of eigenfaces are maximally discriminative of subject groupings (this issue is explored statistically in the discriminant function analysis reported below). Thus, O'Toole et al. (1994) argue that negatively and positively weighted additions of a second eigenface to a first eigenface - produced in their analysis of approximately 100 faces - appears to capture male and female facial information, respectively. Figure 7.6 presents face images constructed as combinations of the first three eigenfaces of Figure 7.5. The combinations appear from visual inspection to resemble prototypic black male and white female faces.



Composites are equally weighted combinations. The black 'patches' on the faces may be an artefact of stroboscopic illumination.

Figure 7.6 Composite images formed from low dimensional eigenfaces.

Procedure

Rating distinctiveness and typicality

Four arrays of 10 face images were constructed, in several distinct sequences,[285] and printed on a Hewlett Packard laser printer at 600 dpi. Each of 41 subjects rated two arrays of images (20 images in total) for distinctiveness, on a 15 point scale. Subjects were instructed as follows: "Imagine that you were to encounter the face in a crowd of people in supermarket mall: how easily would this face 'stick out' in such a crowd? A very distinctive face would 'stick out' to a considerable degree, but a less distinctive face would not 'stick out' as much." (These instructions are similar to those used in several studies by Bruce & Valentine - see Valentine & Bruce, 1986c, for example). Ratings were obtained for a total of 40 faces. Subjects for this rating task, and for all other tasks reported here, were second year university students. They were asked to complete the task during a lecture period.

In addition, each of another 38 subjects rated two arrays of images (20 images in total) for typicality, on a 15 point scale. Subjects were instructed as follows: "Imagine that you were to encounter the face in a crowd of people in a supermarket mall: how likely is it that you will meet someone like this, or someone closely resembling this person? If this is a likely event, then the face is typical". This instruction is similar to one used by Valentine & Bruce (1986c).

Ranking similarity: target arrays

Three arrays of 10 face images were constructed, on the basis of decreasing PCA similarity to a target face in the original set of 62. The Euclidean distances separating the faces are represented on the horizontal axes in Figure 7.10. Target faces were arbitrarily chosen to represent white female, white male, and black male groups. The three arrays were combined in various sequences, and orders, for purposes of counterbalancing, and each of 28 subjects was asked to rank each member of the three arrays in respect of similarity to the target face in each array.

Ranking similarity: pairings of images

Nine arrays of twenty facial images were constructed by selecting images from the original set at random. Each of 111 subjects was asked to create similarity pairings of the images in one of these arrays by i) choosing the most similar pair of faces; ii) the next most similar pair of faces, and so on, until all ten possible (exclusive) pairings had been effected. In all of these tasks, subjects were given a booklet containing the arrays, with instructions, and were asked to complete the tasks during a 20 minute period at the beginning of a lecture.

---

[285] The distinct sequences were created for counterbalancing purposes - that is, to avoid order effects. As in study 1, sequences did not systematically affect ratings, and results will be presented without reference to the sequences.

Results

Results generally showed some support for the prospect of using PC-derived spatial distances as measures of facial similarity. In addition, individual principal components derived from a PC analysis of a set of facial images appear capable of discriminating gross physical differences manifested across race, sex and age groups, and may also be predictive of rated distinctiveness and typicality.

Cluster Analysis

Similarity scores were derived for each possible face pairing from the principal component analysis, and entered into a similarity matrix. Matrix entries, M_{ij}, thus corresponded to the similarity score between the i^{th} and j^{th} faces. This matrix was submitted to hierarchical cluster analysis, with average linkage, using SPSS™ software. The analysis appears to have clustered faces on meaningful physical dimensions. Figure 7.7 attempts to summarize the cluster analysis graphically, depicting 10 clusters identified in the analysis, constituting 41 of the 62 faces. In particular, male and female faces appear to have been efficiently separated (clusters 1, 2, 5, 7 and 8), as have black and white faces (clusters 1, 2, 3, 6, 7, 8, 9 and 10).

Figure 7.7 Clusters of face images identified in the cluster analysis

The classification presented as Figure 7.7 is, of course, only exploratory: the algorithm driving the classification starts by identifying one cluster (all the faces), and proceeds to the point where there are

as many cluster as there are faces (Everitt & Dunn, 1991) - stopping points in between depend on the interpretability of the solution. The stopping point in the present analysis was chosen because it accommodates most of the faces in the image set (41 of 62).

Discrimination of race, sex and age

Principal components derived from the PCA of the set of 62 faces were entered into a stepwise discriminant function analysis as independent variables, with race, sex and age as dependent, or classificatory variables. (Age, and race were categorised for these purposes, as shown in Table 7.4). Discriminant functions, in each of these cases, provided accurate classifications of the images.

In the case of sex, a discriminant function using 10 of the principal components correctly classified 95% of the cases ($\chi^2 = 63.7$; df=10; $p<0.0001$). In the case of race, a discriminant function using 5 of the principal components correctly classified 90% of the cases ($\chi^2 = 95.1$; df=10; $p<0.0001$). In the case of age, a discriminant function using 3 of the principal components correctly classified 82% of the cases ($\chi^2 = 40.9$; df=9; $p<0.0001$). Summary results of the discriminant analyses are reported in Table 7.6

a) Summary of discriminant function analyses

Classification function	Eigenvalue	% variance explained	Canonical correlation	Wilks λ	χ^2	df	p <
Race	3.83	97.5	0.89	0.19	95.1	10	0.0001
Sex	2.18	100	0.82	0.31	63.7	10	0.0001
Age_1	.51	60.7	0.58	0.49	40.9	9	0.001
Age_2	.21	25.0	0.42	0.74	17.33	4	0.001
Age_3	.12	14.2	0.33	0.90	6.44	1	0.01

In the case of Race, there were three categories, and two discriminant functions. The second function was dropped, since it exhibited little classificatory power ($\chi^2 = 5.3$; df=4; $p>0.25$). In the case of Sex, there were two categories, and only one discriminant function. In the case of Age, there were four categories, and three discriminant functions. All three functions contributed significantly to the classificatory power of the analysis, and classificatory results are based on the combination.

b) Discriminant functions

Classification function	Discriminant functions (C's are principal components)	Weights of components in the discriminant functions
Race	C1, C13, C18, C2, C5	-.62, .53, .43, .82, -.69
Sex	C1, C10, C14, C22, C3, C33. C42, C5, C7, C9	.60, .41, -.42, -.47, .73, .36, -.48, -.36, -.77, .50
Age_1	C31, C5, C8	.27, .76, .72
Age_2	C31, C5, C8	.76, .12, -.65
Age_3	C31, C5, C8	.72, -.44, .55

c) Classification tables

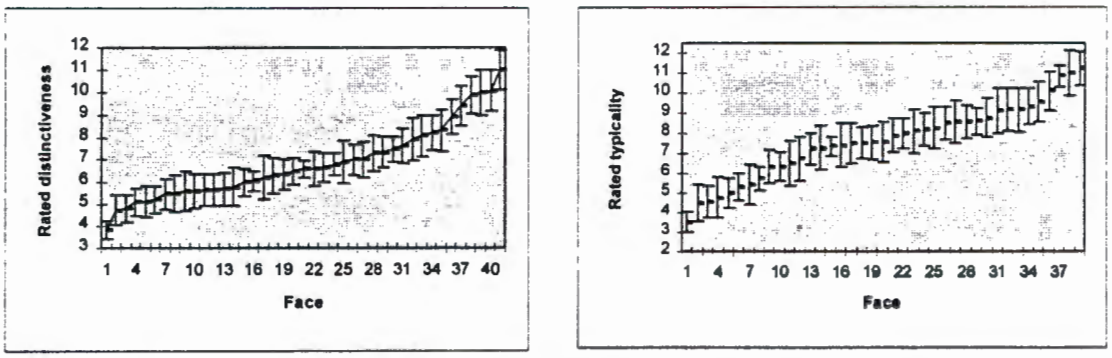
	Race					Sex				Age				
	W	B	C	Tot		F	M	Tot		20	30	40	50	Tot
W	39	0	2	**41**	F	39	2	**41**	20	44	0	0	1	**45**
B	0	7	1	**8**	M	2	19	**21**	30	3	3	0	0	**6**
C	2	1	10	**13**					40	4	0	3	0	7
									50	3	0	0	1	**4**
% Correct classification			*90*				*94*					*83*		

w = white; b=black; c=coloured; f=female; m=male; Tot=Total number of cases. Entries in tables are frequencies, and represent the cross tabulation of actual and predicted group membership - diagonal entries are correct classifications.

Table 7.6 Summary results of the discriminant analysis

Distinctiveness and typicality

Subjects reliably rated faces as differentially distinctive and typical, when ratings were examined in average across the faces, but there was a considerable degree of inter-subject variation in these ratings. Figure 7.8 shows the ratings of distinctiveness and typicality for 41 and 39 faces respectively: standard error bars depict the inter-subject variation.



31 subjects rated distinctiveness, and 37 subjects rated typicality. Both attributes were rated on a 15 point scale. 'Error bars' are constituted by extending a line 1 standard error above the mean, and 1 standard error below the mean. Faces are arranged in order, from least- to most- distinctive, and least- to most- typical.

Figure 7.8 Ratings of distinctiveness and typicality.

Correlations of mean ratings of distinctiveness and typicality with spatial distances of faces from the multivariate mean of the principal component space were negligible ($r = 0.13$ and $r = 0.04$, respectively). It is doubtful that higher-order components in such a space are statistically stable, but I nevertheless investigated the relation between distinctiveness and typicality ratings and principal components, using multiple linear regression. Multiple regression models regressing individual

principal components on these ratings were fairly successful. In the case of distinctiveness, a stepwise regression algorithm identified a model consisting of 10 principal components, with R^2=0.88 ($F = 21.8$; $df = 10, 30$; $p < 0.0001$). In the case of typicality, a stepwise regression algorithm identified a model consisting of 3 principal components, with R^2=0.33 ($F = 5.6$; $df = 3, 34$; $p < 0.003$). The results of the regression analysis are presented in simplified form as Table 7.7

Distinctiveness

Variables in the Equation

Variable	B	St err	β	T	p<
C13	-4.67	1.26	-0.24	-3.71	0.01
C22	7.53	1.39	0.35	5.42	0.01
C23	4.12	1.55	0.17	2.66	0.01
C25	-4.52	1.79	-0.18	-2.53	0.02
C33	-4.74	1.96	-0.16	-2.42	0.02
C41	12.75	2.07	0.40	6.15	0.01
C46	9.50	2.55	0.25	3.73	0.01
C5	4.09	0.75	0.37	5.47	0.01

$R^2 = 0.88$; $R^2adj. = 0.84$

Analysis of Variance

	DF	SS	MS	F	p<
Regression	10	93.39	9.34	21.8	0.001
Residual	30	12.81	0.43		

Typicality

Variables in the Equation

Variable	B	St Err	β	T	p<
C15	-8.21	3.03	-0.38	-2.71	0.01
C20	-8.32	3.64	-0.32	-2.28	0.03
C24	7.04	3.32	0.30	2.12	0.04
Constant	7.57	0.25	30.28	0.00	

$R^2 = 0.33$; $R^2adj. = 0.27$

Analysis of Variance

	DF	SS	MS	F	p<
Regression	3	39.37	13.12	5.58	0.001
Residual	34	79.97	2.35		

Note: C's are principal components. The stepwise procedure was controlled by setting the probability for inclusion to p = 0.06, and for exclusion to p = 0.1

Table 7.7 Summary of regression of principal components on rated distinctiveness and typicality.

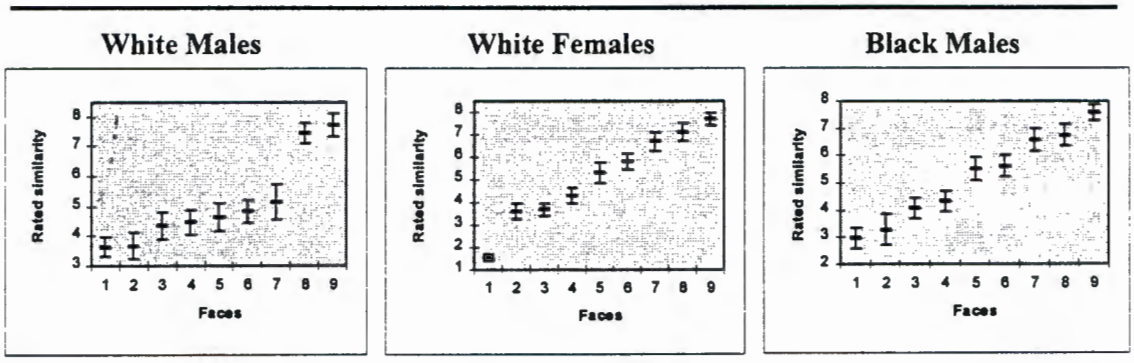
Relation between rated distinctiveness and rated typicality

Although common language usage suggests that (facial) distinctiveness and typicality are antonyms, this is not reflected in their empirical relation. In recent research, Vokey and Read (1988; 1992) report evidence which suggests that rated distinctiveness and typicality are partially independent, and they invoke the notion of 'context-free familiarity' as explanation. In the present research, the relation between rated distinctiveness and typicality was of moderate strength. Ratings of the distinctiveness and typicality of 33 and 37 faces respectively, were provided by independent subjects, as detailed in **Procedure** above, and 29 of these faces were common across stimulus sets. For these 27 faces, mean rated distinctiveness and mean rated typicality were calculated, and the correlation determined across faces, which was statistically significant, albeit of moderate strength ($r = -0.52$; $df = 27$; $p < 0.006$).

Rating similarity: arrays

Subjects reliably ranked faces as differentially similar from target faces, when rankings were examined in average across the faces, but there was a considerable degree of inter-subject variation in these rankings (see Figure 7.9). The differentiation between faces shown here is in marked contrast to

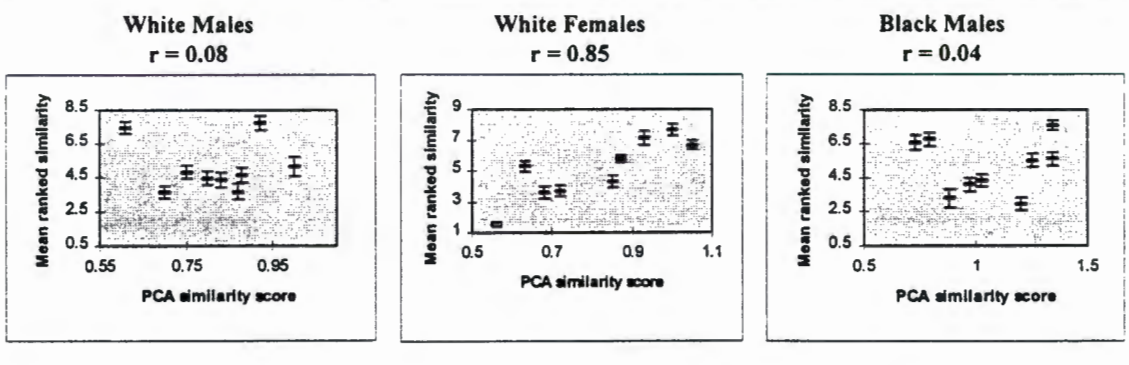
the lack of differentiation shown in Study 1, which is probably consequent on the careful selection of faces for the present arrays.



Error bars extend 1 standard error above and below mean rankings. 28 subjects provided rankings for each array. Faces are arranged in order of mean ranked similarity to targets.

Figure 7.9 Ratings of similarity to target faces in three arrays of ten faces.

In one of the three arrays in which subjects ranked faces in a similarity relation to a target face, the correlation between mean similarity ranking and the ranking derived from the PCA was high, but in the other two the correlation was low. In the case of the array containing the white female target, the correlation was 0.85 ($df = 7$; $p<0.01$); in the case of the array containing the black male target, the correlation was 0.04; and in the case of the array containing the white male target, the correlation was 0.08. Figure 7.10 depicts scatterplots of relations for each of the arrays.



Error bars are lines extended one standard error above and below the mean. 28 subjects provided similarity rankings for each array.

Figure 7.10 Relations between similarity rankings and PCA similarity measures, in three arrays.

Ranking similarity: pairings

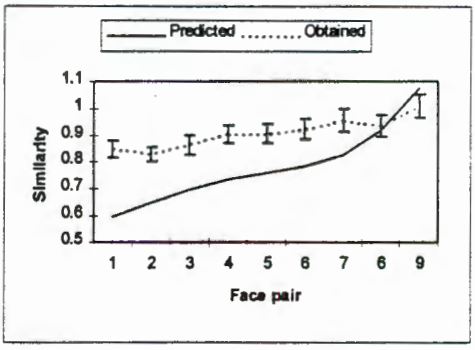
Subjects were required in this task to choose pairs of faces that were most similar to each other, until all possible pairings were exhausted. For each pairing made by subjects a corresponding Euclidean distance was calculated: this was the distance in the principal component space, between the pair of faces selected by the subject. These distances were averaged over subjects so that mean distances were obtained for the 10 exhaustive pairings possible in the task - i.e. each subject produced ten pairings; for each of these pairings a PC-based distance was calculated, and then mean distances were found for each pairing, averaging over subjects. Since the task required subjects to pair the most similar faces in the array, sequentially, and since each array had a determined sequence of closest pairings in terms of the principal component coefficients, it was possible to calculate expected Euclidean distances. Table 7.8 and Figure 7.11 report average observed distances, and expected distances.

Pair no.	1	2	3	4	5	6	7	8	9	10
Expected	0.55	0.60	0.65	0.70	0.73	0.76	0.78	0.83	0.92	1.08
Obtained	0.84	0.85	0.83	0.86	0.90	0.91	0.92	0.96	0.93	1.01
St. dev.	0.15	0.16	0.14	0.19	0.18	0.19	0.2	0.21	0.21	0.22

Standard deviations are reported for observed scores.

Table 7.8 Average obtained and expected Euclidean distances for the pairings task.

There was a very strong correlation between the expected and obtained distances ($r = 0.94$, $df = 8$; $p<0.0001$). The size of this correlation is slightly misleading, since the obtained distances were averaged over 111 subjects, and there was considerable variability between subjects. A better indication of the strength of the relation may be the effect size calculated from an appropriate Analysis of Variance



Error bars are extended one standard error above and below the mean.

Figure 7.11 Relation between observed and predicted Euclidean distance in the face pairing task

Accordingly, a repeated measures single factor Analysis of Variance was conducted across pairings, taking observed Euclidean distance of each face pairing made by subjects as the dependent variable, and ordinal sequence (which had ten levels) as the independent variable. The omnibus test in this analysis was not of much interest, since the PCA measure of facial similarity provided fairly precise predictions of differences between levels of the independent variable. Figure 7.11 shows the predictions, and it is clear from visual inspection that the relation between ordinal sequence and predicted distance is nearly linear.[286] The analysis of variance was therefore conducted by partitioning sums of squares with a set of orthogonal polynomials, which allows one to estimate effects due to linear, quadratic, cubic and other higher order terms. The analysis was implemented under the MANOVA procedure in SPSS, and a summary is reported in Table 7.9

Effect	SS Effect	SS Error	MS Effect	MS Error	F	p <	η^2
Linear	2.54	3.22	2.54	0.03	74.13	0.0001	0.44
Quadratic	0.04	3.82	0.04	0.04	1.04	0.311	0.01
All other*	0.31	-	-	-	-	-	0.11

* i.e. Trends in the range x^3 to x^9. Significance tests are only reported for linear and quadratic terms

Table 7.9 Polynomial trend analysis for results of the face pairing task

It is clear from the trend analysis that the linear effect provides a good fit to the observed pairings data: the effect size was substantial ($\eta^2 = 0.44$), and also statistically significant ($F = 74.13$; $df = 1,94$; $p<0.0001$). Since the performance expected in terms of the PCA similarity measure was very close to constituting a linear relation, the observed pairings data corresponds very well to that predicted by the PCA measure.

Similarity measures derived from low dimensional component spaces

One of the chief benefits of principal component analysis is its ability to reduce a large set of correlated variables to a small set of orthogonal components, which reproduce the variables with relatively little error. The theoretical development of the PCA based measure of similarity outlined in an earlier chapter did not address this issue: the goal of the measure is to capture similarity information, and not necessarily to develop a data-efficient representational scheme. Indeed, all similarity measures used thus far are determined by the full set of principal components, since the full set reproduces the set of face images without any error. There are problems with this approach, though, which led me to explore similarity scores derived from low dimensional component spaces.

[286] It is important to note that the predicted distances are averages, since multiple arrays were used in the pairing task, and each array has a unique pairing sequence. Although the relation depicted in Figure 7.11 appears to deviate slightly from linearity, it is unlikely that fitting non-linear models will be of much use, given the instability of the distance estimates.

In the first place, although the full set of components (say k) perfectly reproduces the original variables, this will be true of any k linearly independent vectors of appropriate length.[287] The components could possess little facial information themselves and yet reproduce a set of face images. In the second place, higher order components will be highly unstable, and will depend almost exclusively on the particular image set they are derived from. Lower order components will be much more stable, and more likely to possess the ability to reproduce images not in the original image set.[288] Ultimately, if the measure is to be of some use, similarity scores for particular faces will need to be generated from a base set of eigenfaces, and not from endless, repeated principal component analyses.

In order to explore the use of a lower dimensional similarity score, the principal component analysis reported above, under **Analysis of stimuli**, was repeated for the images in Study 2, except that the statistical software forced an analysis with just five components.[289] The component solution generated in this fashion accounted for approximately 75% of the variance across face images. Similarity scores derived from the solution were compared to those generated by the full set of principal components, for each of the three arrays used in the similarity rating task (discussed above). Figure 7.12 presents the comparison.

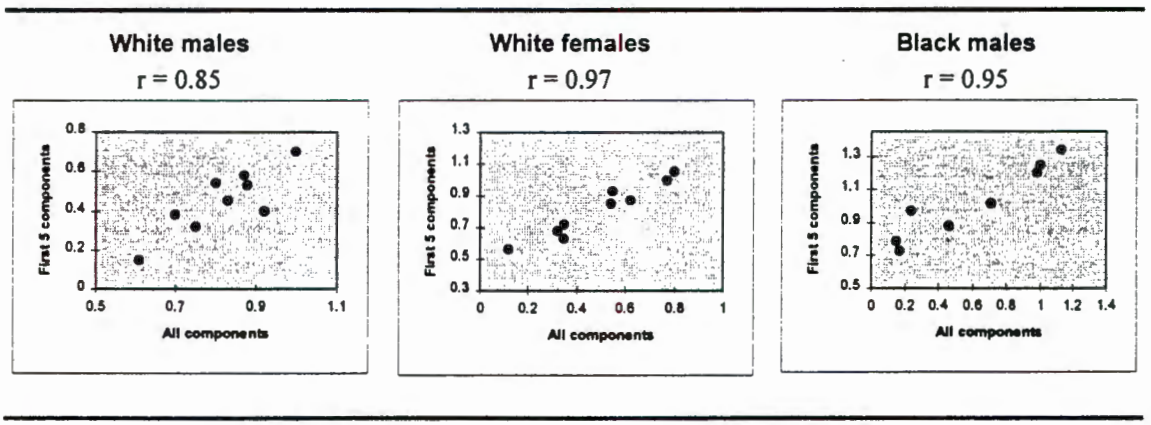


Figure 7.12 Relation between similarity scores derived from the full set of eigenfaces, and scores derived from the set containing the first five eigenfaces.

It is clear that there is a very strong relation between the similarity scores, although the strength of the relation varies quite substantially across the three array types (it is weakest for the array containing white males). This may be due to the under-representation of the white male group in the image set.

---

[287] This is a well known mathematical result. See Lang (1987).

[288] The approximate standard error of eigenvalues (for large n) is given by Flury & Riedwyl (1988) as $s(l_h) = l_h\sqrt{\left(\frac{2}{n-1}\right)}$, where l_h = the h^{th} eigenvalue.

[289] The decision to use five components was based on the size of the eigenvalues, and on an examination of the scree plot (as discussed earlier, on page 169 of this chapter).

Discussion of Study 2, and General Discussion

The results of Study 2 provide support for the claim that PCA-based measures of facial similarity and distinctiveness capture important facial information, and that this information corresponds in reasonable degree to human perceptions and judgements of facial similarity and distinctiveness. Several aspects of the study and the evidence it produced are of some concern, though: in particular, there are issues about i) the practical implementation of the PCA measure, and ii) the enduring problem of inter-subject variability. These will be addressed under several headings, below, alongside a consideration of changes to effect in the research strategy followed in Studies 1 and 2.

Discrimination of 'gross' differences in physical characteristics

The PCA measure of similarity is able, at relatively high levels of accuracy, to discriminate groups of markedly dissimilar faces. Thus, relatively simple component-based discriminant functions distinguished faces of different sex and race with considerable, but imperfect accuracy. Similarly, a simple function was able to discriminate age categories, but with less success.

The degree of inaccuracy associated with the discriminant functions, and its implications, are difficult to evaluate. In the first place, classification tables - such as those presented earlier in Table 7.6 - are known to overestimate the accuracy of the classification (Everitt & Dunn, 1991), and it is useful to examine the degree of accuracy with some form of sample subdivision. This is not attempted in the present chapter, but the issue is taken up in a study to be reported in Chapter 8. In the second place, there is an argument to be made for upper limits on the accuracy of the present discriminant functions: since human judges are rarely able to achieve more than about 95% accurate classification of sex (Bruce et al., 1993), this is perhaps the maximum precision we should expect from discriminant functions that attempt the same.

At any rate, the classificatory ability of the PC analysis is an indication that it is capturing facial information relevant to human perception of facial similarity.

Typicality and distinctiveness

Facial distinctiveness and typicality are key variables in face recognition research and theory: in Chapter 5 I reviewed much of this work, and suggested that distance of a facial image from the origin of a component space may capture facial distinctiveness. However, the empirical investigation reported in this chapter showed that such a measure is a poor indicator of rated distinctiveness and typicality. Nevertheless, principal component analysis does seem to capture important distinctiveness information: component-based regression equations proved highly successful at modelling rated

distinctiveness, but proved less successful in the case of rated typicality. In both cases, few of the early components were selected by the stepwise algorithm for inclusion in the final equation. Thus, even as the success of the modelling suggests that the PCA approach may be a fruitful avenue to pursue in the measurement of facial distinctiveness, it also raises problems. It appears desirable to base PCA similarity (and distinctiveness) scores on a low dimensional component space (I make this argument on page 178, and later in the discussion below), but it seems that distinctiveness and typicality information is captured by higher order components. Measures based on low dimensional subspaces are therefore likely to lose information about distinctiveness and typicality of faces.

Nevertheless, it may be too early to give up on the potential of a PCA based distance measure as an index of facial distinctiveness: the overwhelming problem with results from studies 1 and 2 is that the principal component analysis was based on a comparatively small image set (k=62), and this set itself was disproportionately constituted by young, white females. In Study 3, a much larger and better stratified image set is subjected to PC analysis, and indices of facial distinctiveness are again investigated.

Similarity measures based on lower-dimensional spaces

The measure of facial similarity developed at the end of Chapter 5 assumed that the similarity measure should be calculated with respect to the full component solution i.e. by determining Euclidean distances over all k of the component coefficients in a k-component solution. The full solution perfectly reconstructs all faces in the image set, and a similarity measure based on the full solution would lose no facial information. However, this approach makes the measure highly dependent on the specific solution obtained in the PC analysis, and analyses based on different image sets may produce quite different similarity measures. Furthermore, high order coefficients have very small eigenvalues associated with them, are consequently highly unstable, and it is questionable that they have any statistical meaning (in the sense of constituting variation not attributable to mere sampling fluctuation). If the measure is to have practical utility, it should be possible to find a similarity score for two faces by projecting them into a reasonably stable component space.

Lower-dimensional spaces are more likely to be stable and relatively generalizable, and are a promising solution to this problem: in both studies, results showed that similarity measures based on greatly reduced component spaces corresponded very well to measures derived from the full solution.

The similarity measure and perceived facial similarity

The strength of the relation between mean rated similarity and Euclidean distances derived from the PCA approach is promising, particularly for the pairings task. However, there are several troubling

features. In the first place, two of the arrays in the array ranking task produced results which exhibited virtually no relation between perceived similarity and Euclidean distances. This may be due to the particular constitution of the image set: 66% of the set was drawn from a white female university population, and comparatively few from other groupings in the same population. This is telling in the present case: there was a strong correspondence between rated similarity and the PCA measure in the case of the white female array, and absent relations in the case of the other two arrays, which were constituted by faces drawn from poorly represented segments of the image set. The arrays which failed to show a relation were constituted by faces underrepresented in the sampling plan, and the components that emerged in the analysis may have been inadequate to the task of representing *general* properties of these groups of faces.

Also troubling is the great degree of inter-subject variation in facial similarity ratings: indeed, there is little to model if subjects themselves cannot agree on similarity judgements! Earlier, I mentioned the near-complete absence of discussion in the literature on this issue: although many face recognition researchers have investigated facial similarity, the fact that human similarity judgements vary so greatly has not been addressed. This phenomenon requires careful examination, since most practical applications of the similarity metric assume reasonable degrees of subject equivalence - and I daresay, South African courts assume such equivalence, otherwise the key notion that identification parades should consist of physically similar persons is impracticable.

Reconstructing the similarity task

It appears possible to shed light on the problem of inter-subject variation by constructing alternate similarity rating tasks. Thus, in Study 2, the 'pairings task' revealed a strong relation between the PCA based similarity measure and rated similarity, whereas the array rating task produced results that were equivocal. Since one of the primary objectives in the development of the similarity measure is to provide an independent means of evaluating the fairness of identification parade, it may be useful to explore the relation between the PCA based measure and indices of 'parade fairness' (which were discussed at some length in Chapters 4 and 6). This is done in a later chapter.

It is important, though, to consider possible explanations for the inter-subject variation, rather than merely generating *ad hoc* operationalizations of perceived similarity. This will allow us to think of corrective strategies.

The most bothersome explanation from the point of view of both legal and psychological work on identification parades is that people are inherently inconsistent in perceptions of similarity: that is, perceptions of similarity at time 1 differ substantially from those at time 2. Common sense suggests

that this explanation is false,[290] and it can be investigated with comparative ease, using test-retest techniques.

Another bothersome explanation - fortunately, with a more tractable outcome - is that similarity judgements are made on multiple dimensions, and different subjects tend to base their judgement on different dimensions (or, equivalently, weight the same dimensions differentially). This explanation was considered in the discussion of Study 1, and it appears that there is some evidence for it. It may be useful to explore the basis of similarity judgements in order to obtain further and clearer confirmation. Ways of doing this are explored later in the thesis.

The third, and final explanation I wish to consider here is that the range of perceived similarity is extremely limited (for all subjects). Gross subject differences are perceived (between sex and race groups, for example), but lesser differences are not. Consequently, similarity judgements of targets who are not clearly distinct will be unstable and will show large inter-subject variation. This does not necessarily spell trouble for identification parade practice: in Chapter 4 I reviewed research conducted by Luus & Wells (1991), in which they argue that parades should exhibit 'propitious heterogeneity'. If this explanation turns out to be valid, we will need to investigate the relation of such heterogeneity to measures of parade fairness.

The implication for the PCA measure of similarity is that it might need 'calibration' in terms of this relation: that is, we will need to know what degree of similarity on the PCA measure is predictive of lineups that have poor functional and effective sizes.

Limitations of the image set

The greatest limitation of studies 1 and 2 is the size and constitution of the image sets used there. The set was small, and highly disproportionate in respect of sex, race and age. This constraint is likely to have influenced the similarity (and distinctiveness) measures in several ways. In the first place, the principal component analysis (which sustains the measure) is likely to have produced components that are quite unstable. This means that the components (and therefore, the similarity scores) will have limited generalizability, and that it is probably unwise to apply the similarity metric to faces that were not in the original set. In the second place, since the image set was disproportionately composed of white, female faces, it is probable that the components provide a poor representation of other groups of faces. Similarity scores based on these components will be inaccurate, since they assume that components capture population information. Accordingly, in the further empirical work discussed in

[290] Facial similarity is part of the everyday perceptual world: it is common to remark on the facial similarity of twins, siblings, and families. Stable perceptions surely underlie this talk.

later chapters, a much larger set of face images was collected and submitted to principal component analysis.

Chapter 8

Further empirical tests of the similarity measure

Introduction

In the previous chapter, I reported two empirical studies. The first of these was a preliminary empirical investigation of the facial similarity measure, and the second showed that the measure has promise. The investigations reported there, however, were hampered by several methodological weaknesses, which are addressed in the studies reported in this chapter. In the first place, whereas the earlier investigations used a relatively small image set, which was homogenous with respect to gross subject characteristics (e.g. race and sex), the similarity measure in the present work was derived from a relatively large, heterogeneous image set (278 faces). Other methodological issues addressed in this second set of studies include i) the nature of instructions given to subjects in face rating tasks, and ii) the test-retest reliability of ratings of facial similarity.

A central intention of the research reported in this chapter was to repeat the similarity rating tasks used in the research of Chapter 7. The results there were mixed, and it was not clear whether this was due to the limited image set, the high inter-subject variation, the nature of the rating task, or the measure itself. The results from the re-administration (and modification) of the task are reported here, and they point again to the usefulness of the similarity measure.

The similarity measure is also examined in other substantive ways. In a previous chapter, the similarity of two faces was defined as the distance between them in a principal component space, which is formed by subjecting digitized images of faces to principal component analysis. In component analyses reported earlier in the thesis, frontal photographs of faces constituted the images. However, it is clear that there are many possible viewing angles, and in this chapter I report work which investigates the generality of the similarity measure across two different views of faces, and also explores the possibility of 'combining' views to give a better measure. Finally, I turn to two questions sustaining much of this thesis: can the measure be used as a gauge of the fairness of identification parades?, and what is the relationship between facial similarity and identification accuracy in identification parades? I attempt to answer these questions in two experiments. In the first experiment, similarity of targets in an array was manipulated, and the effect of this manipulation on standard indices of parade fairness was assessed. In the second experiment, subjects were given photographs of people allegedly observed in a particular location, and later tested for their ability to

identify them in a photo-parade. Manipulations in this experiment included varying the similarity of parade members.

Collection and analysis of facial images

The research reported in the previous chapter used a set of images which proved limited in several respects. In particular, the set was relatively small (62 images), and was disproportionately constituted by photographs of young white women. I argued in that chapter that a much larger - and more representative - corpus of images is needed to study perceptions and ratings of facial similarity. The first task for the present research was therefore to collect such a corpus.

Several possibilities were explored regarding possible sources of facial images; I opted finally for a strategy which involved setting up a photographic stall in supermarket malls, and offering passers-by a free photograph in exchange for permission to use this photograph for research purposes. This approach seemed likely to attract a diverse collection of people, and to ensure heterogeneity with respect to gross physical characteristics. (It would be preferable, of course, to construct a sampling frame which ensures representativeness, but it is difficult even to imagine how such a scheme could be constructed in the case of facial characteristics!)

Letters were sent to the management committees of twenty supermarket malls in the Cape Peninsula, requesting permission to operate a photographic stall in the manner described above. Four committees responded positively, and photographs were accordingly collected over a two week period. Photographic subjects were recruited by i) placing posters in conspicuous places, and ii) directly propositioning passers-by. A total of 278 people agreed to pose for photographs. Their characteristics are summarised in Table 8.1, and the (frontal) images are printed as Appendix D.

Age		**Sex**		**Race**[291]	
Category	N	Category	N	Category	N
16 - 19	22	Male	148	Black	14
20 - 29	122	Female	130	Coloured	139
30 - 39	74			White	121
40 - 49	40				
> 50	20				

Table 8.1 Subject characteristics of the 278 facial images collected in supermarket malls.

[291] Race groups reported here are based on those defined in the (now defunct) South African Population Registration Act. They should not be taken to indicate distinct genetic or physiognomic populations, although the groups do differ considerably in physical appearance.

It was not possible to standardize ambient lighting in the different locations. Instead, subjects were positioned in front of a grey matte screen, and a stroboscopic flash unit provided a direct source of light. Two 35mm format cameras were used to take photographs (a Canon EOS 500, and Canon EOS 100), at a focal length of approximately 80mm. Exposure was controlled by the automatic TTL metering system of each camera. Subjects were asked to adopt a neutral expression (as in a 'passport' photograph), and to look straight ahead at the camera. A second photograph was then taken, with subjects adopting a ¾ profile position.

Photographic film was later developed by the author, contact printed on Ilford 'Pearl' photographic paper, and digitally scanned at 300 dpi on a Hewlett Packard IIx flatbed grey-image scanner.

In research reported in the previous chapter, digital images were edited to remove jewellery and other extraneous costume. This process is laborious, and will surely be unattractive to people who wish to use the similarity measure at a practical level. Accordingly, I explored the effect of leaving jewellery and costume intact on derived similarity scores.

Similarity scores and extraneous costume

Twenty facial images were randomly selected from the larger set of 278 frontal images. Images were standardized according to the scheme outlined in Chapter 7, to a size of 120 x 150 = 18 000 pixels: that is, all images were scaled to equalise inter-ocular distance, and positioning of the left and right pupils.[292] These images were submitted along with the remaining 258 images to principal component analysis, and complete component coefficients were computed for each image. An inter-correlation matrix was determined by correlating the co-ordinates of each of the 20 images with the co-ordinates of every other image in the set of 20 (each image is defined by the PCA as a set of component co-ordinates). The twenty images were then digitally edited to remove jewellery and other extraneous adornments, and the process was repeated as for the unmodified images. Two correlation matrices were thus formed, one containing inter-correlations of unmodified images, and the other containing inter-correlations of modified images. These matrices were differenced, and the resulting absolute differences appear to be small, albeit non-zero (see Table 8.2, where the average absolute difference is reported for each image).

[292] In the case of profile views, faces were equalised on chin-nose distance, and positioning of the left pupil. The principal component analysis of profile views was in other respects identical to that pursued for frontal views.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
0.1	0.07	0.04	0.06	0.06	0.09	0.09	0.08	0.1	0.03
F11	F12	F13	F14	F15	F16	F17	F18	F19	F20
0.04	0.08	0.1	0.04	0.05	0.04	0.06	0.09	0.11	0.11

Median difference = 0.065. F1 = face image 1, F2 = face image 2, and so on.

Entries were calculated as follows: each face was correlated with every other face, in both modified and unmodified forms. Absolute differences were computed for each of these correlations, and averaged columnwise (i.e. for each face). Thus, the entry 0.1 under F1 indicates that the average absolute difference in inter-correlations between F1 and all other faces, across modified and unmodified forms, is 0.1

Table 8.2 Absolute differences in inter-correlations across modified and unmodified images.

The effect of using unmodified images on similarity scores is thus probably quite small, and further analyses were computed using only these images. (Further investigations may of course show that the small median difference reported above is non-trivial).

Principal component analysis and the similarity metric

In Chapter 7 I noted that it is probably inefficient to derive the similarity measure from the full set of component coefficients generated by principal component analysis. High order components typically resolve negligible amounts of variance, and including them may hardly alter similarity scores (see Chapter 7, page 178 onward). The problem is to identify an adequate number of components for the component space. This problem is typically dealt with in multivariate statistical practice by using a number of 'rules of thumb', viz. Kaiser's 'eigenvalues > 1' rule, and Catell's 'scree plot'. In the present case, both the scree plot and Kaiser's rule proved of little help. Figure 8.1 presents two scree plots - in the first of these, the scale the eigenvalues take makes the plot impossible to interpret, and in the second, there is no discernible 'elbow'.

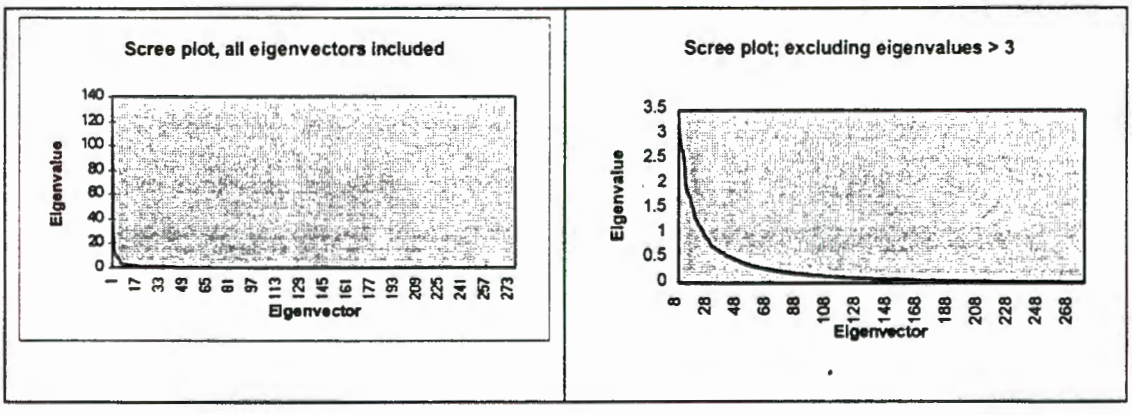


Figure 8.1 Scree plots of eigenvalues from the PCA of 278 frontal facial images

Table 8.3, on the other hand, shows that excluding all eigenvalues < 1 would lead to a 24 component solution which leaves 15% of the solution unresolved, and is surely unsatisfactory. An alternative rule suggested by Joliffe (1972), which leads to excluding eigenvalues < 0.7, does little better.

No.	Value	Cum% var.	No.	Value	Cum% var.	No.	Value	Cum% var.	No.	Value	Cum% var.	No.	Value	Cum% var.	No.	Value	Cum% var.	No.	Value	Cum% var.	No.	Value	Cum% var.
1	123.94	44.60	36	0.64	89.00	71	0.26	94.0	106	0.14	96.40	141	0.09	97.80	176	0.06	98.70	211	0.04	99.30	248	0.03	99.80
2	40.40	59.10	37	0.61	89.20	72	0.25	94.10	107	0.14	96.50	142	0.09	97.90	177	0.06	98.80	212	0.04	99.40	249	0.03	99.80
3	13.48	64.00	38	0.61	89.40	73	0.25	94.20	108	0.14	96.50	143	0.09	97.90	178	0.06	98.80	213	0.04	99.40	250	0.03	99.80
4	9.66	67.40	39	0.60	89.60	74	0.24	94.30	109	0.14	96.60	144	0.08	97.90	179	0.06	98.80	214	0.04	99.40	251	0.03	99.80
5	7.77	70.20	40	0.55	89.80	75	0.24	94.40	110	0.13	96.60	145	0.08	98.00	180	0.06	98.80	215	0.04	99.40	252	0.02	99.80
6	6.77	72.70	41	0.55	90.00	76	0.24	94.50	111	0.13	96.70	146	0.08	98.00	181	0.05	98.80	216	0.04	99.40	253	0.02	99.80
7	4.39	74.30	42	0.54	90.20	77	0.23	94.60	112	0.13	96.70	147	0.08	98.00	182	0.05	98.90	217	0.04	99.40	254	0.02	99.80
8	3.20	75.40	43	0.52	90.40	78	0.22	94.60	113	0.13	96.80	148	0.08	98.00	183	0.05	98.90	218	0.04	99.40	255	0.02	99.80
9	2.92	76.50	44	0.50	90.60	79	0.22	94.70	114	0.13	96.80	149	0.08	98.10	184	0.05	98.90	219	0.04	99.50	256	0.02	99.80
10	2.71	77.40	45	0.50	90.80	80	0.22	94.80	115	0.12	96.90	150	0.08	98.10	185	0.05	98.90	220	0.04	99.50	257	0.02	99.90
11	2.68	78.40	46	0.47	90.90	81	0.21	94.90	116	0.12	96.90	151	0.08	98.10	186	0.05	98.90	221	0.04	99.50	258	0.02	99.90
12	2.44	79.30	47	0.47	91.10	82	0.21	94.90	117	0.12	97.00	152	0.08	98.20	187	0.05	98.90	222	0.04	99.50	259	0.02	99.90
13	1.99	80.00	48	0.45	91.30	83	0.21	95.00	118	0.12	97.00	153	0.08	98.20	188	0.05	99.00	223	0.04	99.50	260	0.02	99.90
14	1.82	80.60	49	0.43	91.40	84	0.20	95.10	119	0.12	97.00	154	0.08	98.20	189	0.05	99.00	224	0.03	99.50	261	0.02	99.90
15	1.79	81.30	50	0.42	91.60	85	0.20	95.20	120	0.12	97.10	155	0.07	98.20	190	0.05	99.00	225	0.03	99.50	262	0.02	99.90
16	1.72	81.90	51	0.41	91.70	86	0.20	95.20	121	0.11	97.10	156	0.07	98.30	191	0.05	99.00	226	0.03	99.50	263	0.02	99.90
17	1.54	82.50	52	0.40	91.90	87	0.19	95.30	122	0.11	97.20	157	0.07	98.30	192	0.05	99.00	227	0.03	99.60	264	0.02	99.90
18	1.52	83.00	53	0.39	92.00	88	0.19	95.40	123	0.11	97.20	158	0.07	98.30	193	0.05	99.10	228	0.03	99.60	265	0.02	99.90
19	1.32	83.50	54	0.38	92.10	89	0.19	95.40	124	0.11	97.20	159	0.07	98.30	194	0.05	99.10	229	0.03	99.60	266	0.02	99.90
20	1.27	83.90	55	0.37	92.30	90	0.18	95.50	125	0.11	97.30	160	0.07	98.40	195	0.05	99.10	230	0.03	99.60	267	0.02	99.90
21	1.22	84.40	56	0.37	92.40	91	0.18	95.60	126	0.11	97.30	161	0.07	98.40	196	0.05	99.10	231	0.03	99.60	268	0.02	99.90
22	1.16	84.80	57	0.34	92.50	92	0.18	95.60	127	0.10	97.40	162	0.07	98.40	197	0.05	99.10	232	0.03	99.60	269	0.02	99.90
23	1.12	85.20	58	0.34	92.60	93	0.17	95.70	128	0.10	97.40	163	0.07	98.40	198	0.05	99.10	233	0.03	99.60	270	0.02	100
24	1.07	85.60	59	0.33	92.80	94	0.17	95.80	129	0.10	97.40	164	0.07	98.50	199	0.05	99.20	234	0.03	99.60	271	0.02	100
25	0.99	85.90	60	0.33	92.90	95	0.17	95.80	130	0.10	97.50	165	0.07	98.50	200	0.05	99.20	235	0.03	99.60	272	0.02	100
26	0.93	86.30	61	0.33	93.00	96	0.16	95.90	131	0.10	97.50	166	0.07	98.50	201	0.05	99.20	236	0.03	99.70	273	0.02	100
27	0.91	86.60	62	0.32	93.10	97	0.16	95.90	132	0.10	97.50	167	0.07	98.50	202	0.04	99.20	237	0.03	99.70	274	0.02	100
28	0.87	86.90	63	0.31	93.20	98	0.16	96.00	133	0.10	97.60	168	0.06	98.60	203	0.04	99.20	238	0.03	99.70	275	0.02	100
29	0.83	87.20	64	0.30	93.30	99	0.16	96.10	134	0.10	97.60	169	0.06	98.60	204	0.04	99.20	239	0.03	99.70	276	0.02	100
30	0.76	87.50	65	0.30	93.40	100	0.16	96.10	135	0.09	97.60	170	0.06	98.60	205	0.04	99.30	240	0.03	99.70	277	0.01	100
31	0.74	87.80	66	0.29	93.50	101	0.15	96.20	136	0.09	97.70	171	0.06	98.60	206	0.04	99.30	241	0.03	99.70	278	0.01	100
32	0.71	88.00	67	0.29	93.70	102	0.15	96.20	137	0.09	97.70	172	0.06	98.60	207	0.04	99.30	242	0.03	99.70			
33	0.70	88.30	68	0.28	93.80	103	0.15	96.30	138	0.09	97.70	173	0.06	98.70	208	0.04	99.30	243	0.03	99.70			
34	0.67	88.50	69	0.28	93.90	104	0.15	96.30	139	0.09	97.80	174	0.06	98.70	209	0.04	99.30	244	0.03	99.70			
35	0.65	88.70	70	0.26	93.90	105	0.15	96.40	140	0.09	97.80	175	0.06	98.70	210	0.04	99.30	245	0.03	99.70			

No. = Number of eigenvector; Value = eigenvalue of corresponding eigenvector; Cum. var. % = cumulative percentage variance resolved by current and all preceding eigenvectors. Eigenvalues are rounded to two places.

Table 8.3 Eigenvalues and cumulative resolved variance from the PCA of 278 frontal views

Accordingly, I pursued a different way of thinking about the number of eigenvectors on which to base the component space. Since the goal of PCA in the present case is to constitute a space in which faces are accurately represented by common, orthogonal components, I examined reconstructions of faces based on increasing numbers of eigenvectors. Reconstructions of ten randomly chosen faces based on the first 24, 50, 75, 100 and 140 eigenvectors are shown as Figure 8.2.

Reconstructions are weighted linear combinations of eigenvectors (or principal components). The weights are the coefficients of principal components produced by PCA

Figure 8.2 Reconstructions of 10 face images with increasingly large sets of eigenvectors

Figure 8.2 clearly shows that the set of eigenvectors with associated eigenvalues > 1 (i.e. the first 24 eigenvectors) do not satisfactorily reconstruct facial images. Indeed, from visual inspection of the reconstructions, it appears that at least 100 eigenvectors are required to produce facial images which approximate their original form. This supports a claim O'Toole et al. (1994) make, namely that higher order components carry important information of identity. Perhaps the measure of facial similarity does not need to be sustained by a representational basis that is precise with respect to identity, but the reconstructions based on 24 and 50 images are quite imprecise, and this does not invite confidence in highly reduced component solutions. All similarity scores computed in the present research are thus based on solutions with 100 eigenvectors.

The assumption of linearity

In Chapter 5, I briefly discussed alternate ways of standardizing face images prior to PCA. One alternative was devised by Craw & Cameron (1991), with the intention of maintaining several of the axioms of linear systems. I disregarded this approach earlier, since a study by Hancock et al. (1994)

appeared to show that there is little difference between the standard and modified approaches. In the analysis of the large set of images collected for the present research, I decided to return to this issue.

The problem Craw & Cameron identified with the standard approach to PCA of face images, is that it breaks the axioms for vector addition and scalar multiplication. The linear space that has its basis in the eigenvectors produced by the PCA of face images must meet the following axioms (at least):[293]

Scalar multiplication	*Vector addition*
If **v** is in V, then $a\mathbf{v}$ is in V for all a in $\mathbb{R}$	If **u** and **v** are in V, then $\mathbf{u} + \mathbf{v}$ is in V

where V is a vector space, **u**, **v** are vectors, a is a scalar, $\mathbb{R}$ is the set of real numbers.

PCA produces a basis for representing faces in terms of common components, and this basis works on the assumption that the axioms above are satisfied: each face in the image set is represented as the weighted combination of eigenvectors i.e. the scalar multiplication and addition of eigenvectors. The problem is that certain combinations of faces in the image set are not incontrovertibly members of 'face space'. The standardization scheme suggested by Craw & Cameron corrects the 'blurring' problem (see Chapter 5), but it is not clear to me that it satisfies the axioms outlined above. This is because combinations of the basis vectors of 'face space' produce faces that are quite unlike faces, even when a corrective scheme is used like that suggested by Cameron and Craw. Figure 8.3 shows i) the first ten eigenfaces produced by the PCA of the frontal views of the image set of 278 faces; and ii) combinations of eigenfaces weighted according to the covariance structure of the component coefficients.

E1 = eigenface 1, E2 = eigenface 2, and so on. Fic1 = fictitious face 1, Fic2 = fictitious face2, and so on.

Figure 8.3 The first 10 eigenfaces from the PCA, and 10 'random'[294] combinations of the eigenfaces.

[293] See Lang, (1987), or any first level linear algebra text.

None of the combinations depicted above are anything like real faces, although all have shape and form closer to faces than to any other objects. There appear to be two important implications for present purposes. The first is that the kinds of differences the combinations above have in relation to real faces are not restricted to shape, and this means that the scheme suggested by Craw & Cameron will not ensure that PC-based representational schemes meet the axioms of linear systems. An important qualification in this respect is that it is difficult to know what degree of correspondence between combinations and real faces is acceptable - the images in the figure above resemble real faces, even if they are clearly not faces. The second implication of breaking the axioms is that the measure of facial similarity proposed in this thesis may rest on tenuous ground. The similarity between two faces is defined as the Euclidean distance between the image vectors representing the faces in component space, but this distance is meaningful only if the space is linear. Again, this depends on how strictly we evaluate the resemblance between combinations of basis vectors and real faces. If we are prepared to accept a component space which is 'facelike', rather than a space constituted only by real faces, the consequences will be less serious.

Study 3: Ratings of similarity and distinctiveness

Similarity

In the first of the empirical studies reported in this chapter, I returned to the rating tasks which provided mixed support for the measure of facial similarity in Study 2. In these tasks, subjects are asked to rate the similarity of a number of people to a designated 'target', and ratings are correlated with distances of the faces in component space. One of the tasks in Study 2 yielded results which showed a strong relation between rated similarity and spatial distances, but the remaining two did not. However, there were methodological uncertainties in Study 2, which I discussed in Chapter 7. I attempted to address these uncertainties in Study 3.

In the first place, the image set in Study 2 was severely limited by both size and homogeneity. The corrections in this respect have been outlined at some length above. In the second place, the rating task in the earlier study required subjects to *rank* stimuli in terms of similarity to a target. This may have narrowed the range of differences in perceived similarity available as responses to subjects, and the task was altered in Study 3 to allow subjects to *rate* similarity on a continuous scale. In the third place, the task of explicitly rating facial similarity is an unusual one. Although it is likely that most humans frequently make implicit judgements of similarity, to be asked to do so in a formal task is

[294] A weight was selected at random for the first component (within the range produced by the PCA for the set of 278 faces on the first component), and weights for a further 99 components were generated from the covariances between the first component and the next 99 components in the PCA of 278 frontal faces.

unusual, and invites unusual behaviour. This may make subjects adopt a mechanistic and simplified approach (e.g. using one facial feature as the basis for the judgement). Accordingly, subjects were given instructions in Study 3 which attempted to structure their judgement.

Other issues were addressed in Study 3, which were not specifically identified as methodological weaknesses in Study 2. Whereas Study 2 did not examine the effect of facial distinctiveness on similarity ratings, Study 3 varied the distinctiveness of targets in order to assess separate and interactive effects of distinctiveness. In the face pairings task reported in Chapter 7, it appeared that subjects were able to distinguish gross differences in similarity, but were not able to make fine distinctions. Accordingly, tasks were structured in Study 3 so that arrays were constituted by members who differed 'continuously' or 'discontinuously' on the similarity distance measure. That is, arrays were either constituted by several groups of members who differed in substantial 'steps' of similarity from the target (i.e. 'discontinuously'), or by members who differed in small, evenly spaced 'steps' of similarity from the target (i.e. 'continuously'). If subjects are able only to make gross distinctions, their judgements should match the spatial distance measure better in 'discontinuous' arrays than in 'continuous' arrays.

Subjects

Subjects were 76 Psychology 1 students at the University of Cape Town. They participated in the study during a lecture.

Materials

Twelve array tasks were constructed in a similar manner to those used in Study 2. Ten face images were printed on paper in array form, and one of these was designated as the target. Three arrays were joined with a covering page of instructions into an experimental booklet. The arrays varied according to the design of the study, presented below. A sample booklet is reproduced as Appendix E. Instructions issued to subjects were altered from those used in Study 2, in line with the rationale outlined earlier:

> Over the page you will find three collections of faces. The collections are marked "A", "B", and "C", respectively. In each collection you will notice that each face has a number below it, except one, which is called the 'target' face. I would like you to compare the numbered faces to the target face - how similar in facial appearance are they to the target?
>
> In order to arrive at the rating of similarity, use any facial quality you think relevant. You may also wish to take the following into consideration:
>
> Hair: are the length, colour, and texture similar?
> Hairline: is the hairline equally high or low on the forehead?
> Face shape: do the faces have the same shape (e.g. round, thin, angular)?
> Noses: are the noses similar in size and shape?
> Mouths: are the mouths equal in size? are the lips equally full?
> Skin texture and colour: is this similar?
> Chins: are these the same shape and size?
>
> Please use a scale with the extreme values shown below. You can assign any number between 0 and 10.

0 _____ 5 _____ 10

Not at all similar Highly similar

Please indicate for each face, how similar it is to the target face. Do this by writing the value you have chosen next to the number of the face. Repeat this process for each of the three collections, "A", "B" and "C".

Design

The principal data of interest in Study 3 were the similarity ratings subjects made in array tasks. However, these arrays were constructed to vary in terms of the distinctiveness of the target (low, moderate, high), and in terms of the graded similarity of the members of the parade (discontinuous, continuous). The similarity and distinctiveness of parade members was determined from the PC-based distance measure. In addition, each of the six arrays created in this way was formed in two distinct sequences, for counterbalancing purposes. The design is schematised in Figure 8.4

		Graded similarity	
		Discontinuous	Continuous
Distinctiveness	High	34	37
	Moderate	35	37
	Low	38	35

Cell entries are numbers of subject responses, and not numbers of subject per se, since subjects completed three tasks. (The 'distinctiveness' factor was a repeated measure).

Figure 8.4 Design for Study 3 - similarity rating tasks.

Procedure

Subjects were asked at the beginning of a lecture to participate 'in a study on face perception and recognition', after which experimental booklets were distributed. These booklets had been arranged in a randomized order prior to distribution, and the cover page gave subjects all further necessary instructions. The booklets themselves variously contained arrays which were instances of conditions and sequences described in 'Materials' and 'Design' above.

Results

Results of the similarity rating task provided strong positive support for the notion that the PC-based similarity metric corresponds in reasonable degree to ratings of similarity made by human subjects. The facial distinctiveness of targets and the graded similarity of arrays affected ratings of similarity, but not so greatly as to negate the correspondence between the spatial distance measure and subject ratings.

The dependent variable for analyses of distinctiveness and graded similarity effects was the correlation between individual subject ratings of similarity and the spatial distance measure. Each correlation was transformed to approximate standard normal deviate form for purposes of analysis,

using the Fisher transformation.[295] Three separate between-subjects one way analyses of variance were conducted in order to assess the effect of the graded similarity manipulation: one for each level of the distinctiveness factor.[296] These analyses showed that the graded similarity manipulation depressed rating correspondence for arrays with targets of low facial distinctiveness only. A one way repeated measures analysis of variance was then conducted to assess the distinctiveness manipulation, and this showed that arrays with targets of low distinctiveness depressed rating correspondence - but Figure 8.5 appears to show that this is due to the graded similarity manipulation. Table 8.4 and Figure 8.5 report the results of the analyses.

1 Way ANOVAs (graded similarity manipulation)

		df Effect	MS Effect	df Error	MS Error	F	p <
Distinctiveness	High	1	0.12	72	0.10	1.26	0.26
	Moderate	1	0.05	72	0.10	0.56	0.45
	Low	1	0.57	72	0.13	4.47	0.04

Repeated Measure ANOVA (target distinctiveness manipulation)

df Effect	MS Effect	df Error	MS Error	F	p <
2	0.53	146	0.09	6.11	0.002

Table 8.4 Analysis of variance tables for graded similarity and distinctiveness manipulations.

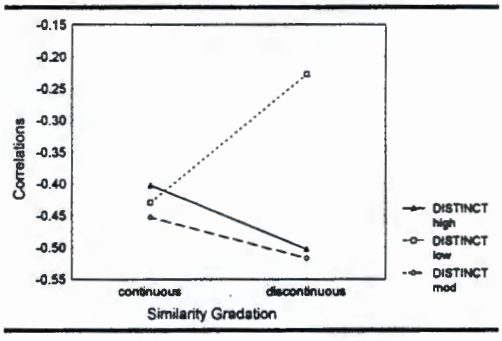


Figure 8.5 The effect of the distinctiveness and similarity gradation manipulations on correspondence between subject ratings and the spatial distance measure of similarity.

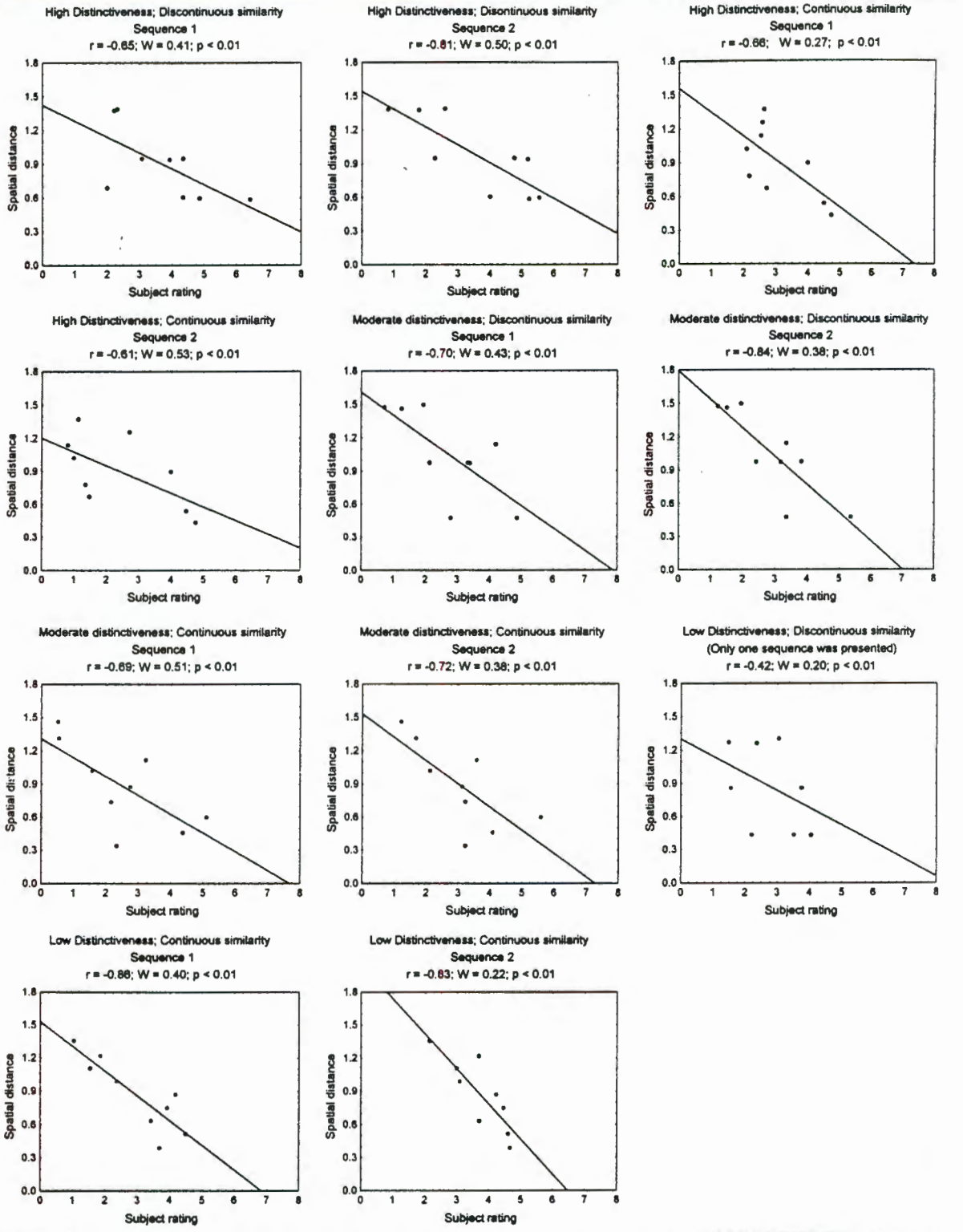
In order to assess the correspondence between subject ratings of similarity and the spatial distance measure in an analogous manner to that used in Chapter 7, average similarity ratings were computed across subjects, and correlated with spatial distances, for each of the twelve arrays. Scatterplots are shown in Figure 8.6, and annotated to show correlation coefficients and Kendall concordance coefficients.[297]

---

[295] This is a recommended practice when treating correlations as data. See Howell (1992).

[296] The full study design consisted of one repeated measures factor (facial distinctiveness) and one between subjects factor (graded similarity). However, subjects were not assigned consistently to conditions in the graded similarity factor - i.e. a subject could have been assigned a high distinctiveness array with discontinuous graded similarity, and then a moderate distinctiveness array with continuous graded similarity, and so on. This made the model specification for a two way analysis of variance rather cumbersome, and so I opted for the simpler approach to the analysis, as outlined here.

[297] Ratings were transformed to ranks in order to compute Kendall coefficients.



Points are mean ratings. All p values are calculated for Kendall's coefficient (W). W is calculated for subject ratings only, and reflects the degree of agreement among subjects. Negative relationships are expected, since the rating and spatial distance scales take reverse directions.

Figure 8.6 Relations between subject ratings of similarity, and the PC measure of facial similarity.

Figure 8.6 shows that there is a strong correspondence between subject ratings of facial similarity and the spatial distance measure of facial similarity. In each of the eleven array tasks, the relationship is in the expected direction, and the (absolute) correlation is always greater than 0.40 in size. The median absolute correlation is 0.70, which is strong. The consistency of both size and direction are

convincing demonstrations that the spatial distance measure corresponds in reasonable degree to human judgements of similarity.[298] Nevertheless, it should be remembered that average ratings tend to inflate correlations, and that subjects were far from consistent in their ratings, even with the modifications to the task instructions used in this study. (The inconsistency is shown in the size of the Kendall coefficients: these show that agreement was better than chance expectation, but not complete).

Distinctiveness

I argued in an earlier chapter that the PC-based representational basis offers two facial measures - a measure of facial similarity, and a measure of facial distinctiveness. The similarity of two faces is the distance between the faces in component space, whereas facial distinctiveness of a face is the distance of that face from the origin[299] of the component space. In Chapter 7, I explored the relationship between subject ratings of facial distinctiveness and the PC distinctiveness measure, but found that there was little relation. It proved possible, however, to successfully model ratings of facial distinctiveness by regressing component coefficients on distinctiveness ratings. I returned to this question in Study 3.[300]

A possible explanation for the failure of the distinctiveness measure in Study 2, is that the basis vectors of the space were determined from analysis of a small, homogenous image set. The basis determined in Study 3 may yield different results, since it was determined from a much larger image set. Accordingly, the distinctiveness rating task deployed in Study 2 was administered, with arrays of faces taken from the frontal image set, and ratings were correlated against distinctiveness scores derived from a PCA of this set.

Subjects

Subjects were 26 first year Medical students at the University of Cape Town, attending a year-end course evaluation meeting. They were asked to participate in a face recognition study, and completed rating forms at the beginning of the evaluation meeting.

---

[298] The results of Study 4, which are reported later in the chapter, show a similar, strong correspondence and further bolster this conclusion.

[299] Distance from the multivariate mean is a more practical solution; see footnote 284 in Chapter 7.

[300] The distinctiveness research that is reported here was conducted separately i.e. at a different time, and with different subjects, and is reported under Study 3 merely for convenience.

Materials and Procedure

Three arrays were constructed, consisting of 28, 28 and 26 face images, respectively.[301] These arrays were printed on a Hewlett Packard laser printer at 600 dpi. Ten subjects completed the first array, seven completed the second, and nine completed the third. Subjects rated arrays for distinctiveness, on a 15 point scale. Subjects were instructed in the same way as subjects in Study 2: "Imagine that you were to encounter the face in a crowd of people in supermarket mall: how easily would this face 'stick out' in such a crowd? A very distinctive face would 'stick out' to a considerable degree, but a less distinctive face would not 'stick out' as much." These instructions were printed on the cover page of an experimental booklet, which was distributed to subjects. Ratings were obtained for a total of 82 faces.

A mean distinctiveness rating was determined for each face image, and the resulting 82 mean ratings were correlated with the PC based facial distinctiveness measure. A scatterplot of the relation is reported as Figure 8.7

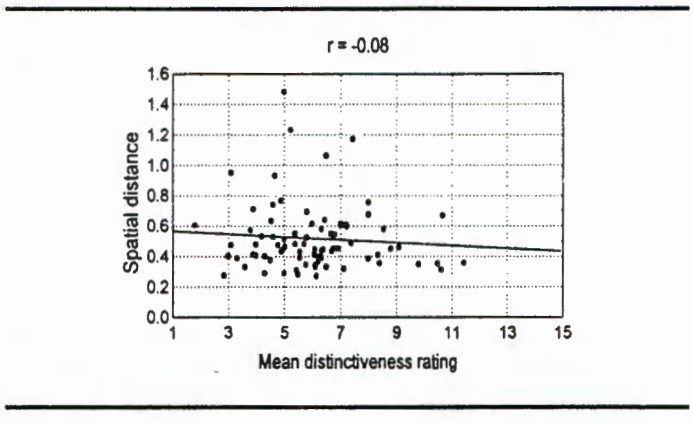


Figure 8.7 Relation between rated distinctiveness and the spatial distance measure of distinctiveness.

It is clear from Figure 8.7 that there is little relation between rated distinctiveness and the PC-based measure. This is in keeping with the finding made in Study 2: basing distinctiveness scores on a larger image set appears to have little effect on this relation. Despite the absence of a relation between PC-based distinctiveness, defined as distance from the multivariate mean, and rated distinctiveness, regression analyses in Study 2 showed that it is possible to use weighted combinations of individual components to model rated distinctiveness. Similar analyses were conducted on data collected in the present study - coefficients and statistics of a model identified by a stepwise regression procedure are reported as Table 8.5.

[301] The distinct sequences were created for counterbalancing purposes - that is, to avoid order effects. As in study 1, sequences did not systematically affect ratings, and results will be presented without reference to the sequences.

Variables in the equation

Var.	B	St. err	β	T	p <	Var.	B	St.err	β	T	p <
C2	-1.10	.37	-0.20	-2.95	0.01	C63	10.57	4.83	0.015	2.18	0.03
C26	5.45	2.42	0.15	2.25	0.02	C64	-11.18	4.47	-0.17	-2.49	0.02
C31	11.73	2.82	0.30	4.14	0.00	C74	-25.84	4.55	-0.42	-5.66	0.01
C32	5.97	2.36	0.17	2.53	0.01	C75	-15.26	4.39	-0.24	-3.47	0.01
C36	10.62	2.76	0.27	3.83	0.01	C78	12.52	4.50	0.19	2.78	0.01
C44	-10.84	2.78	-0.26	-3.89	0.01	C79	-16.54	4.87	-0.23	-3.39	0.01
C57	-13.47	3.61	-0.25	-3.72	0.01	C87	-14.07	5.74	-0.17	-2.45	0.02
C58	18.33	4.30	0.32	4.25	0.01	C98	-22.58	5.41	-0.28	-4.17	0.01
Const.	6.02	0.13		44.44	0.01						

$R^2 = 0.71$; $R^2_{adj.} = 0.64$
Standard error of estimate = 1.13

Analysis of Variance

	DF	SS	MS	F	p <
Regress.	16	206.09	12.88	10.16	0.001
Residual	65	82.39	1.27		

Note: C's are principal components. The stepwise procedure was controlled by setting the probability for inclusion to p = 0.06, and for exclusion to p = 0.1

Table 8.5 Summary of regression of principal components on rated distinctiveness; data from Study 3.

Although the model resolves a reasonably high amount of variance, it should be remembered that stepwise regression procedures capitalize on chance, and frequently produce models that are unstable. Cross-validations of the model were therefore attempted, on randomly selected subsamples of the distinctiveness data set. In the first set of cross-validations, I evaluated the regression equation formed from the variables identified in Table 8.5, on two sub-samples.[302] The resulting equations resolved a satisfactory, and statistically significant, amount of variance (Equation 1: $R^2_{adj.} = 0.74$; $F = 9.12$; $df = 16, 30$; $p < 0.001$; Equation 2: $R^2_{adj.} = 0.65$; $F = 4.88$; $df = 16, 18$; $p < 0.001$). In the second set of cross-validations, I conducted a stepwise regression on a further two randomly selected subsamples (each subsample had 43 subjects). The point of this exercise was to see whether components identified in the two analyses would match i) those identified by the original stepwise procedure, and ii) match across the two subsamples. Variables identified in the three stepwise procedures are reported in Table 8.6

Original	Subsample 1	Subsample 2
C2, C26, C31 C32, C36, C44 C57, C58, C63 C64, C74, C75 C78, C79, C87 C98	C2, C31, C36 C44, C46, C47 C59, C6, C69 C94, C98	C13, C19, C51 C54, C58, C59 C66, C68, C70 C78, C81, C85
$R^2_{adj.} = 0.66$ $F = 10.16$ $p < 0.001$	$R^2_{adj} = 0.80$ $F = 13.4$ $p < 0.001$	$R^2_{adj} = 0.81$ $F = 15.5$ $p < 0.001$

Note: C's are principal components. All three procedures were controlled by setting probability of inclusion in the model to 0.06, and probability of exclusion to 0.1

Table 8.6 Components selected by stepwise regression procedures.

[302] Regression coefficients were re-determined for each of the sub-samples, using statistical software.

The table shows that the three procedures identified sets of components with comparatively little overlap. It is not entirely clear what this means: stepwise regression is notoriously fickle, and one expects variable selection to be a little inconsistent, but it leaves the question of measuring distinctiveness no closer to a solution.[303] This is the last empirical investigation of the spatial distance measure of distinctiveness to be reported in this thesis, so I leave it as an issue for later research, or other researchers.

Study 4: Viewing perspective and facial similarity

In Chapter 5, I pointed to an important problem in much face recognition research, which is the use of single viewing perspectives - typically photographs taken from a frontal perspective. Studies which use single viewing perspectives cannot distinguish face perception and memory from picture perception and memory, and probably yield inflated estimates of recognition ability, in particular. Viewing perspective is also an important consideration for the type of measure of facial similarity proposed in this thesis, viz. a measure based on analysis of a set of images, standardized for viewing perspective. I explored the role of viewing perspective in several ways in Study 4. In the first place, I examined the relation between similarity scores derived from analysis of a set of frontal images, and scores derived from analysis of a set of ¾ profile views. These scores should be strongly correlated, and the absence of a strong correlation would be evidence against the robustness of the PC-based similarity measure. However, what would constitute a 'strong correlation' here is not clear, so I attempted to establish some baseline value by correlating similarity ratings of frontal and profile views made by human subjects. In addition, I explored the relation between the PC-based measure of similarity and subject ratings of similarity, for both frontal and profile views, and combinations of these. Finally, I briefly examined the utility of combining principal component analyses of frontal and profile views.

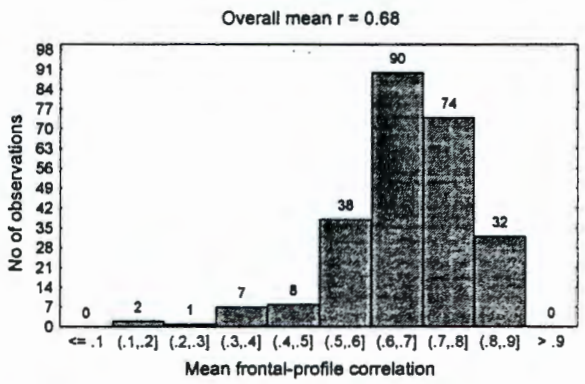
Relation between PC-based similarity scores across viewing perspectives

As described in an earlier section of this chapter, photographs were taken of 278 people from frontal and ¾ profile viewing perspectives, converted to digital images, and submitted to principal component analyses. In the case of the profile views, several of the images were discarded, leaving a total of 257 images.[304] The set of ¾ profile images is reproduced as Appendix F. The principal component analysis of profile images was conducted in the same manner as that for frontal images, as were

[303] It is interesting, however, that almost none of the early components get selected for inclusion in any of the three models.

[304] It proved difficult to take successful photographs from a ¾ profile perspective. Subjects were urged to adopt an appropriate stance, and much effort was made to manoeuvre subject and camera into a suitable relation, but this was not always possible. About 30 profile photographs were discarded. In contrast, only five frontal photographs were discarded. There were 252 images for which both frontal and profile images were available for analysis.

similarity scores, and neither will therefore be described here. Profile and frontal similarity scores were obtained for each face image for which both profile and frontal views were available: that is, the (PC-based) similarity of each face to every other face in the image set was determined, for both frontal and profile image sets. Since the aim was to determine whether the image sets generated equivalent similarity relations, a correlation was calculated between the set of frontal and profile similarity scores, for each face image. The distribution of this correlation is shown in Figure 8.8



Note: in order to calculate the overall mean correlation, individual correlations were Fisher-transformed before computation, and the mean of these transformed correlations was inverse-transformed. Individual correlations represent the relation between frontal and profile similarity scores, calculated for each face image.

Figure 8.8 Distribution of correlations between frontal and profile similarity scores.

The similarity relations are clearly not equivalent across frontal and profile views, but are certainly not insubstantial. The important question here is how strong a relation is acceptable: a perfect relation is improbable, as faces will show differences when viewed from different angles, and this can be expected to attenuate the strength of the relation. On the other hand, it is unlikely that similarity relations will change dramatically with a change of viewing perspective. In order to assess the strength of the relation shown in Figure 8.8, I investigated the relation between similarity ratings made by subjects when shown frontal views, and ratings made when the same faces were shown in ¾ profile view. At the same time, I (again) investigated the relation between rated similarity and the PC-based measure of similarity, and the effect of facial distinctiveness on similarity ratings.

Subjects

Subjects were 90 Psychology 1 students at the University of Cape Town. They participated in the study during a lecture.

Materials

Twelve array tasks were constructed in a similar format to that used in Study 3. Ten face images were printed on paper in array form, and one of these was designated as the target. Two arrays were joined

with a covering page of instructions into an experimental booklet. The arrays varied according to the design of the study, presented below. A sample booklet is reproduced as Appendix G. Instructions issued to subjects were the same as those used in Study 3, as outlined on page 193 above.

Design

The principal data of interest in Study 4 were the similarity ratings subjects made in array tasks, which varied in terms of whether frontal, profile, or both frontal and profile views were presented to subjects. Face images selected for inclusion in arrays varied substantially on the PC-based similarity measure, to strengthen the manipulation.[305] Arrays were also constructed to vary in terms of the (PC-based) distinctiveness of the target (low, and high). In addition, each of the six arrays created in this way was formed in two distinct sequences, for counterbalancing purposes. The design is schematised in Figure 8.9

		Distinctiveness	
		High	Low
View	Frontal (F)	33	32
	Profile (P)	33	22
	F+P	24	22

Cell entries are numbers of subject responses, and not numbers of subject per se, since subjects completed two tasks. (The 'distinctiveness' factor was a repeated measure).

Figure 8.9 Design for Study 4.

Procedure

Subjects were asked at the beginning of a lecture to participate 'in a study on face perception and recognition', after which experimental booklets were distributed. These booklets had been arranged in a randomized order prior to distribution, and the cover page gave subjects all further necessary instructions. The booklets themselves variously contained arrays which were instances of conditions and sequences described in 'Materials' and 'Design' above.

Results

Similarity ratings of frontal, profile, and frontal + profile views were strongly related, and the correspondence between these ratings and spatial distances was again fairly high. These results again

[305] Similarity scores for the combined frontal and profile view were calculated by computing the Euclidean distance of each face from the target face, across coefficients from each of the component solutions (i.e. the component solutions were treated as one solution).

point to the usefulness of the PC-based similarity measure. Also, distinctiveness, and viewing perspective appear to affect the degree of correspondence between subject ratings of similarity, and the PC-based measure of similarity, but it is not clear what this means.

Table 8.7 shows relations between frontal, profile, and frontal + profile ratings, and relations between spatial distances calculated for each viewing perspective from the principal component analysis. In each case where the result for subject ratings are reported, reported correlations involve mean subject ratings, and not individual subject ratings.

a) Inter-correlations of subject ratings of similarity

i) Low distinctive target

		Frontal		Profile		Frontal + Profile	
		Seq1	Seq2	Seq1	Seq2	Seq1	Seq2
Frontal	Seq1	1.00					
	Seq2	0.92	1.00				
Profile	Seq1	0.77	0.83	1.00			
	Seq2	0.58	0.78	0.85	1.00		
F + P	Seq1	0.85	0.94	0.94	0.90	1.00	
	Seq2	0.56	0.58	0.84	0.78	0.74	1.00

ii) High distinctive target

		Frontal		Profile		Frontal + Profile	
		Seq1	Seq2	Seq1	Seq2	Seq1	Seq2
Frontal	Seq1	1.00					
	Seq2	0.57	1.00				
Profile	Seq1	0.38	0.57	1.00			
	Seq2	0.34	0.81	0.83	1.00		
F + P	Seq1	-	-	-	-	-	
	Seq2	0.53	0.71	0.45	0.41	-	-

Note: The wrong member in sequence 1 of the F + P condition, high distinctiveness array, was inadvertently labelled as the 'target', and inter-correlations are therefore not reported for this condition.

b) Inter-correlations of spatial distance similarity scores

i) Low distinctive target

		Frontal		Profile		Frontal + Profile	
		Seq1	Seq2	Seq1	Seq2	Seq1	Seq2
Frontal	Seq1	1.00					
	Seq2	1.00	1.00				
Profile	Seq1	0.86	0.86	1.00			
	Seq2	0.86	0.86	1.00	1.00		
F + P	Seq1	0.95	0.95	0.97	0.97	1.00	
	Seq2	0.95	0.95	0.97	0.97	1.00	1.00

ii) High distinctive target

		Frontal		Profile		Frontal + Profile	
		Seq1	Seq2	Seq1	Seq2	Seq1	Seq2
Frontal	Seq1	1.00					
	Seq2	1.00	1.00				
Profile	Seq1	0.82	1.00	1.00			
	Seq2	0.82	1.00	1.00	1.00		
F + P	Seq1	0.95	0.95	0.95	0.95	1.00	
	Seq2	0.95	0.95	0.95	0.95	1.00	1.00

c) Correlations of similarity ratings and spatial distance similarity scores

i) Low distinctive target

Similarity ratings

		Frontal		Profile		Frontal + Profile	
		Seq1	Seq2	Seq1	Seq2	Seq1	Seq2
Frontal	Seq1	-0.70					
	Seq2		-0.79				
Profile	Seq1			-0.91			
	Seq2				-0.78		
F + P	Seq1					-0.91	
	Seq2						-0.77

Spatial distance

ii) High distinctive target

Similarity ratings

		Frontal		Profile		Frontal + Profile	
		Seq1	Seq2	Seq1	Seq2	Seq1	Seq2
Frontal	Seq1	-0.20					
	Seq2		-0.25				
Profile	Seq1			-0.88			
	Seq2				-0.63		
F + P	Seq1					-	
	Seq2						-0.09

Spatial distance

Table 8.7 Inter-correlations of similarity ratings and spatial distances.

Several things are clear from the table. In the first place, it is clear that inter-correlations of spatial distances across frontal, profile, and combined views are at least as strong as those between subject ratings across the same views. This seems to resolve the question of how strong the relation between

frontal and profile spatial distances should be (i.e. it is strong enough), and is further evidence in support of the utility of the spatial distance measure. In the second place, correlations between subject ratings of similarity and spatial distance estimates of similarity are again high (albeit not uniformly), and in the expected direction.

Results reported immediately above were based on mean subject ratings. In order to assess effects due to distinctiveness and viewing perspective, I examined individual subject ratings of similarity. The dependent variable for purposes of analysis was the correlation between individual ratings of similarity and the spatial distance measure. Each correlation was transformed to approximate standard normal deviate form (reported mean correlations are inverse transformed from this form). A one way repeated measures analysis of variance was conducted to assess the distinctiveness manipulation, and this showed that arrays with targets of low distinctiveness increased rating correspondence over arrays with targets of high distinctiveness ($F = 29.57$; $df=1,86$; $p < 0.001$). However, this was not uniformly the case: Figure 8.10 shows that there was little difference between profile views of arrays with highly distinctive targets. One way between subjects analyses of variance were conducted to assess differences over viewing perspective, for arrays with high-distinctive and low-distinctive targets. In the former case, there were significant differences between viewing perspectives ($F = 18.08$; $df=2,86$; $p < 0.001$), but this was not true for the latter ($F = 0.75$; $df=2,86$; $p > 0.47$)

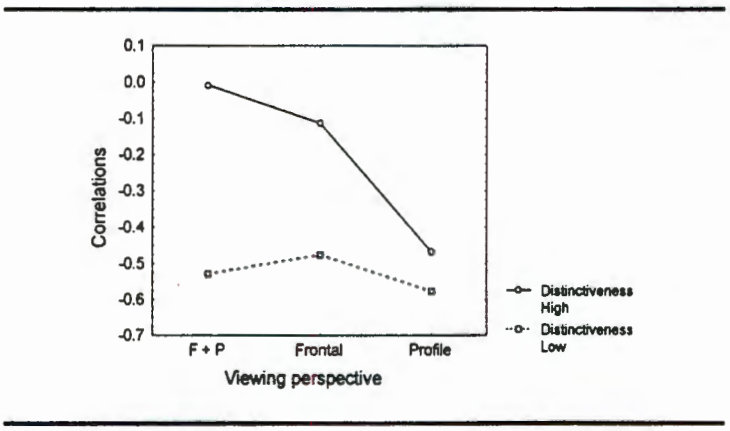


Figure 8.10 Effects of viewing perspective and target distinctiveness on correspondence between subject ratings and the spatial distance measure of similarity.

These results appear to show that particular viewing perspectives decrease the correspondence between subject ratings and the spatial distance measure, but only when targets are of high facial distinctiveness. Several factors make me reluctant to accept these results as they stand: in the first place, they conflict with findings made in Study 3, where high distinctiveness improved the correspondence and low distinctiveness worsened it. In the second place, results from the

distinctiveness rating tasks in Study 3 were that subject ratings of distinctiveness bear little relation to the PC-based measure of distinctiveness: it is thus unclear what the distinctiveness manipulation in this study represents, in terms of perceptions held by subjects.

Study 5: Test-retest reliability of subject ratings of similarity

In Studies 1 and 2, I noted that ratings of similarity show considerable inter-subject variance, and considered a few explanations for this lack of concordance. I will not review that discussion here, except to note that the most bothersome explanation is that people are inherently inconsistent in perceptions of similarity: if that is the case, perceptions of similarity at time 1 will differ substantially from those at time 2. In the present study, I investigated the reliability of similarity ratings over time, using a standard test-retest procedure.

Subjects

Subjects were 21 Psychology 3 students. They participated in a face rating task during the final tutorial of term, and a follow-up rating task some three weeks later.

Materials

Three face rating tasks were prepared for use in the study. Each rating task followed the format used in Study 2: i.e. an array of 10 faces was printed on a sheet of paper, one of which was designated as the 'target' face. Subjects were then asked to rate each face in terms of similarity to the target face. A 100 point scale was used in the present case (as opposed to 10 point scales in the earlier studies). The first of the three arrays was used in the initial rating completed by subjects, and two further arrays were created for administration at the follow-up stage. Each of the additional arrays was created by removing a number of the faces presented in the initial array, replacing them with different faces, and changing the order of presentation of the five original faces. Five faces were removed in one of the arrays, and the other four faces (constituting the original array of nine non-target faces) in the second of the arrays. The initial and follow-up arrays are reproduced as Appendix H.

Procedure

Subjects were approached during the last tutorial session of term, and asked to participate in a face perception and recognition study. Two groups of subjects, attending different tutorial sessions, participated in the study. Each subject was given a rating task, which had the necessary instructions appended as a cover page. Subjects were asked to provide names and student numbers, but were not

told to what purpose they would be put. After a period of two weeks, each subject (bar two)[306] was contacted by post, and asked to complete a second rating task. Rating tasks were attached to letters requesting subjects to participate: there were two versions of the follow up rating task, and these were randomly distributed amongst subjects. Twelve of the nineteen subjects who were asked to participate in the follow-up stage of Study 5 submitted completed rating tasks.

Results

Test-retest reliability of similarity ratings proved to be fairly good, although this varied substantially across subjects.

Correlations for ratings of the five (or four) faces which appeared in both initial and follow-up tasks were computed for each subject Table 8.8 shows the correlation coefficient for each subject, and also reports the median correlation, and correlation over mean ratings.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Correlation	0.00	0.13	0.23	0.51	0.58	0.68	0.73	0.86	0.92	0.94	0.99	1.00

Median correlation = 0.7 Correlation between mean ratings (i.e. over subjects) of initial and follow-up arrays = 0.94

Table 8.8 Correlations between initial and follow-up ratings, per subject.

The correlations are acceptably high, especially when computed over subjects. However, it is prudent to bear in mind the methodological problems usually associated with test-retest designs: in particular, subjects may show demand effects, and attempt to recreate ratings from memory, rather than from a fresh scrutiny of the task. The three week period between test and re-test is some security against this threat, as is the insertion of five (or four) new faces, and the re-arrangement of remaining faces within arrays. On the other hand, the use of only five (or four) faces across tasks, renders the estimate of the test-retest correlation coefficient a little unreliable: the median estimate of 0.7 may be too high, or too low.

Discussion of Studies 3, 4 and 5

There were central and peripheral aims in Studies 3, 4 and 5. The central aim was to rigorously test the facial similarity measure derived from principal component analysis of facial images, on face

[306] Postal details for two subjects were missing from university records, and these subjects were not asked to complete the follow-up task.

rating tasks. This was effected with i) a larger image set, ii) arrays which exhibited greater variation of facial similarity, and iii) tasks which structured subject judgements. I also investigated the generality of the PC-based similarity measure across viewing perspectives. Peripheral aims included testing the facial distinctiveness measure, derived from principal component analysis of face images, and assessing the effect of manipulated distinctiveness on similarity rating tasks.

The tests of the facial similarity measure were largely positive - in both Studies 3 and 4, correlations between mean subject ratings and the PC-based measure were high. There were 24 rating tasks in total across the two studies, of which 17 produced correlations greater than 0.65 in size, and 11 produced correlations greater than 0.75. If one bears in mind that subject ratings of similarity show considerable variation - which means that mean ratings will not be very reliable - this correspondence is good. There were several tasks, however, in which the correlation between subject ratings and the PC based similarity measure was much lower, albeit always in the expected (negative) direction. It is not clear why the results in Studies 3 and 4 show a clear correspondence, where the results of Study 2 were mixed. This could be due to the fact that similarity scores were derived from analysis of a much larger image set: the image set used in Study 3 consisted largely of young, white females, and the representational basis on which the similarity scores rely was probably inaccurate for faces from other populations. Alternatively, it could be due to the fact that arrays were composed of face images that varied more greatly on the similarity measure than in the previous study. Alternatively, again, it could be due to the revised instructions, which attempted to structure subject ratings. There is no easy way to evaluate these explanations on the basis of the data from Studies 3 and 4, and I will not attempt an evaluation here. It is enough for present purposes that the data shows convincingly that the PC based measure corresponds reasonably well to subject ratings of similarity.

The investigation of viewing perspective in Study 4 produced further evidence in favour of the similarity measure. Just as face recognition studies which test subject for their memory of face images run the risk of mistaking picture memory for face memory, so a similarity measure based on just one view of a face runs the risk of mistaking view-similarity for face similarity. Similarity measures taken from analyses of separate views of faces (frontal and ¾ profile) proved to be fairly strongly related. The strength of this relation furthermore appeared quite adequate when compared to the relation of subject similarity ratings across different viewing perspectives.

Clear as the results of the similarity rating tasks were, so results involving the PC-based distinctiveness measure were equivocal. This measure was defined in an earlier chapter as the Euclidean distance from the multivariate mean of component coefficients, but proved unrelated to subject ratings of facial distinctiveness in both Studies 2 and 3. Although stepwise regression procedures showed that it is possible to model subject ratings by weighted composites of component coefficients, models of this sort proved unstable. Different components were identified by the

stepwise procedure for cross-validation subsamples, and this makes a general measure an unlikely solution. (A measure based, for example, on a weighted combination of coefficients of components could be quite useful, but not if the identity of the components is uncertain).

Certain manipulations involving the PC-based distinctiveness measure produced statistically significant effects. In Study 3, correlations between subject ratings and the PC-based similarity measure were smaller for arrays with targets of low distinctiveness, but only when array similarity was structured discontinuously. In contrast, correlations were smaller in Study 4 for arrays with targets of high distinctiveness, but only when images were viewed in frontal or combined frontal-profile orientation. It is difficult to interpret this set of results: they are (partially) contradictory, and since the PC-based distinctiveness measure is not related to subject perceptions of distinctiveness, the meaning of the effects is unclear. This is a pity, since perceived distinctiveness has proved to be an important variable in recent face recognition research (see Chapter 5).

At this stage of the empirical research, it is probably fair to conclude that measures of facial similarity derived from principal component representational bases are sufficiently closely related to subject ratings of similarity to use them as approximations of perceived similarity. However, it is apposite to recognise that perceived similarity can have many operationalizations, and that the rating tasks used in Studies 1 to 5 are merely several instances. In particular, the rationale behind much of the empirical research reported in this thesis was to develop a measure of lineup fairness, on the assumption that perceived similarity of lineup members is an important factor. It is time, therefore, to report research which looks directly at the relation between lineup similarity and lineup fairness, and at the usefulness of the measure of facial similarity in this respect.

Study 6: Facial similarity and mock witness parades.

In Chapters 4 and 6 I discussed a method of evaluating parade fairness, which is commonly used in psycho-legal research. This is the method of the mock witness: mock witnesses are given a description of a suspect, and asked to identify the suspect from an identification parade. Several measures of parade fairness are associated with this method; I evaluated these in Chapter 6 and made certain recommendations regarding their usage and interpretation. In Study 6, I examined the relationship between the measure of facial similarity and measures of parade fairness, using the mock witness method.

Subjects

Subjects were 169 first year medical students at the University of Cape Town. They completed mock witness tasks at the beginning of a year-end course evaluation.

Materials

Three face images were selected from the frontal set of 278 to serve as suspects in mock witness tasks. The images were selected so as to vary on the PC-based facial distinctiveness measure; to wit, as highly distinctive, moderately distinctive and of low distinctiveness. This was achieved by selecting images at random from the top 5%, middle 5%, and bottom 5% of the distribution of distinctiveness scores. The images are shown as Figure 8.11

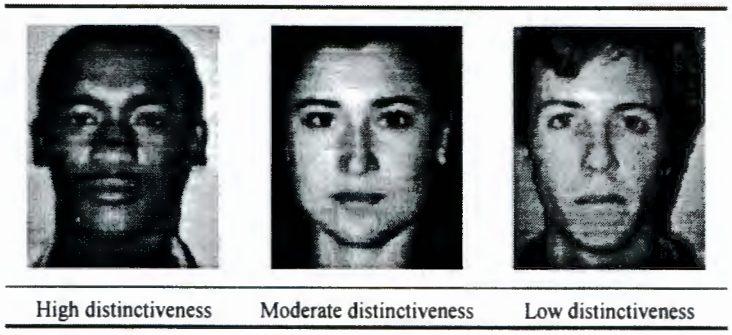


Figure 8.11 Suspects selected for the mock witness tasks.

These images were then given to three raters, in different orders, to get verbal descriptions for the mock witness tasks. Raters were shown each image for twenty seconds, and after presentation of the image were instructed as follows:

> Now, please describe the person you have just seen, without turning back to the previous page. This description should be highly accurate - other people should be able to identify the person on the basis of the description alone - but write what you can remember, even if you think that the description is not accurate enough.

Written descriptions were then carefully examined by the author, and combined to form a description of each suspect. The final descriptions are shown in Table 8.9

High distinctiveness	Moderate distinctiveness	Low distinctiveness
A coloured or black male, probably in his late 20's or early 30's. He has dark, short, curly hair, and dark eyes; a round face, slightly flattened head, and round chin. He has a close, scanty beard and moustache; a large, flat nose, and full eyebrows. His ears are small.	A young white female with dark, thick, slightly curly hair hanging down her back. She has straight, full eyebrows, and large, dark eyes. She has a round face, with wide, well defined jawbones, a small, pointed chin, and a closed, smallish mouth. She has striking looks.	A slim, white male, probably in his early twenties. He has an angular, thin face, with high cheekbones, a square chin, and fullish lips. He has straight, short hair; and teeth which are slightly bucked.

Table 8.9 Descriptions of suspects used in the mock witness tasks.

Six photo-parades were then created, for each of the suspects. These parades were constructed so as to structure the similarity of parade members in relation to the target. This was achieved by selecting images which were in the first 6^{th}, second 6^{th}... sixth 6^{th} of the distribution of similarity scores, calculated in relation to the suspect. In this way, three sets of six parades, of differing target-member similarity, were constructed, making 18 parades in total. These parades are shown in Appendix I. Three parades - one selected from each of the three sets, at random - were combined, along with a

covering page of instructions, into an experimental booklet. Six of these booklets were created in total. Instructions required subjects to identify suspects on the basis of the descriptions provided them:

> Over the page, you will find three collections of pictures. Each collection is of a number of people. The collections are labeled "A", "B", and "C" respectively.
>
> One of the people in each collection was recently seen in a shopping mall in Cape Town. A description of each of these 'target' people is provided below. Your task is to guess who the person is, on the basis of the description.
>
> Please indicate your guess by circling the number below the person in each collection who you think matches the corresponding description. Since there is a correct answer to the problem, you must choose one person in each of the three collections.

Design

The chief aim of Study 6 was to determine the relation between measures of lineup fairness and the spatial distance measure of facial similarity. Accordingly, subjects were presented with photo-lineups which varied systematically on measured facial similarity: there were six 'degrees' of similarity, defined in terms of the sextiles of the distribution of similarity scores.[307] However, parades were also structured according to the facial distinctiveness of the suspect (see the discussion in Materials, above), resulting in a total of three distinctiveness conditions. Similarity and distinctiveness conditions were cross-combined, producing 18 conditions in total, as shown in Figure 8.12

	Similarity					
	(Most similar)					(Least similar)
	1	*2*	*3*	*4*	*5*	*6*
High	23	32	26	31	26	30
Moderate	31	27	30	26	30	23
Low	30	23	26	31	32	26
Distinctiveness						

Cell entries are numbers of subject responses, and not numbers of subjects *per se*, since subjects completed three tasks. (The 'distinctiveness' factor was a repeated measure). Similarity conditions are numbered 1 - 6, from most to least similar.

Figure 8.12 Design for Study 6.

Procedure

Subjects were addressed at the beginning of a year-end course evaluation meeting, and asked to participate in a study of face recognition and perception. Experimental booklets, which were pre-arranged in random order with respect to manipulations constituting the design of the study, were then

[307] 'Similarity' here refers to the similarity between parade members treated as suspects, and other members of the parade.

distributed to subjects. Subjects completed the tasks independently, at their own pace, after which booklets were collected.

Results

There was a strong relation between measures of lineup fairness and the spatial distance measure of similarity: higher degrees of similarity between suspects and parade members led to greater lineup fairness.

Several dependent measures were formed, since several measures of lineup fairness are currently used in psycho-legal research, and one of the aims of work reported in this thesis is to compare the measures. At the simplest level, a binary dependent variable was created, according to whether subjects had chosen correctly or incorrectly on mock witness tasks. A log-linear analysis on a three way table embodying the design of the study (Similarity X Distinctiveness X Correctness[308]) showed that a model incorporating all main effects, and the following interaction effects - i) similarity x correctness, and ii) distinctiveness x correctness - produced a satisfactory fit to observed frequencies. (L.R. $\chi^2 = 25.4$; df = 20; $p > 0.19$).[309] In other words, it was not necessary to include the three way interaction in the model, nor was it necessary to include the remaining two way interaction (distinctiveness x similarity): the effects of similarity and distinctiveness were independent of each other. Figure 8.13 shows the proportion of accurate identifications per experimental condition, as well as the reciprocal of the proportion, which is the measure of lineup fairness known as 'functional size'.

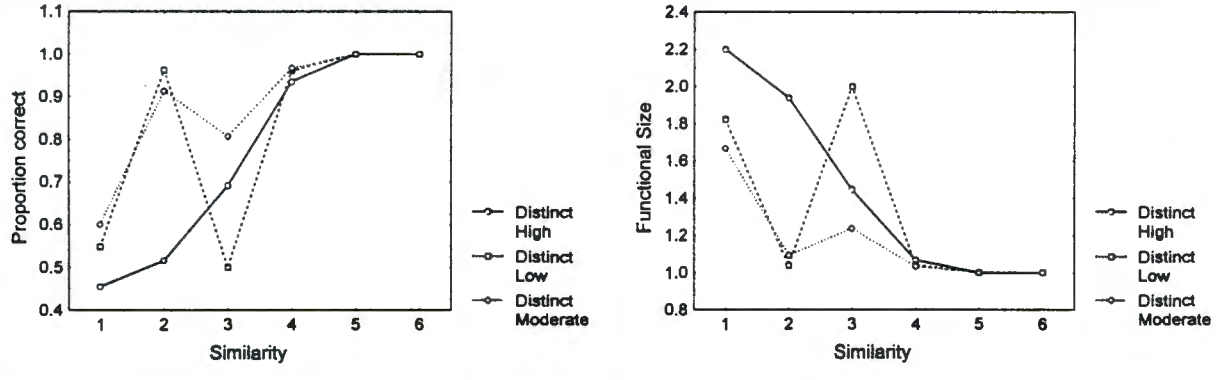


Figure 8.13 Similarity and distinctiveness effects on mock witness accuracy.

[308] 'Correctness' is the binary dependent variable created by coding responses as correct or incorrect.

[309] This analysis assumes independence of observations, which is not the case, since the 'distinctiveness' factor was a repeated measure. I use much the same kind of analysis in Study 7, where I more fully discuss the implications of breaking the assumption.

The similarity effect is evident in Figure 8.13, but the distinctiveness effect is difficult to interpret. The fluctuation of responses to parades with the low distinctiveness target hinders the interpretation, particularly in the similarity = 3 condition: without it, low and moderate distinctiveness appear to decrease identification accuracy, more or less uniformly, in relation to high distinctiveness. What is clear, though, is that increasing similarity of parade members to the suspect reduces the likelihood that the identity of the suspect can be guessed by witnesses armed with only a brief verbal description.

Alternate measures of lineup fairness were then computed. These included the measures known as 'Effective size'[310], and 'E'. Since these measures were discussed at considerable length in Chapters 4 and 6, I will not discuss their formulation or rationale here. Correlations between measures of lineup fairness and a pseudo variable, representing facial similarity,[311] were computed over the 18 conditions of Study 6, and are reported in Table 8.10

	Similarity	Esize (1)	Esize (2)	E
Similarity	-			
Effective size (1)	-0.72	-		
Effective size (2)	-0.78	0.97	-	
E	-0.77	0.96	0.94	-
Functional size	-0.73	0.89	0.86	0.97

Table 8.10 Correlations between measures of lineup fairness, and facial similarity.

Correlations between alternate measures of lineup fairness are very strong: in particular, 'E', which is recommended over measures of effective size in Chapter 6, since it accommodates inferential quantitative methods, is almost perfectly correlated with both formulations of effective size. Correlations of all measures with lineup similarity are strong, and in the expected direction.

Three of the measures of lineup fairness tabulated above are intended to produce an estimate of the number of 'feasible', or 'good' foils present in a lineup.[312] Their absolute sizes therefore have meaning, and are reported in Table 8.11, for each of the 18 lineups used in Study 6. Estimates of functional size are included for comparison.

[310] Two measures were computed here; i) using the formula originally proposed by Malpass, and ii) using the revised formula, as set out in Chapter 6.

[311] The pseudo variable was constituted by the numbers 1 - 6: i.e. the ordinal sequence of similarity conditions in the study.

[312] That is, 'effective size (1)', 'effective size (2)' and 'E'. Functional size gives an estimate of the bias against the suspect in a lineup, although Wells et al. (1978) originally claimed that it serves to estimate number of good foils

High distinctiveness					Moderate distinctiveness					Low distinctiveness				
Simil.	Esize (1)	Esize (2)	E	Fsize	Simil.	Esize (1)	Esize (2)	E	Fsize	Simil.	Esize (1)	Esize (2)	E	Fsize
1	2.55	3.09	2.60	2.20	1	2.35	3.00	2.47	1.82	1	2.27	2.53	2.11	1.67
2	2.97	3.55	2.74	1.94	2	1.07	1.30	1.08	1.04	2	1.26	1.70	1.19	1.10
3	2.85	3.23	1.97	1.44	3	2.10	2.27	2.13	2.00	3	1.58	2.54	1.49	1.24
4	1.13	1.52	1.14	1.07	4	1.08	1.31	1.08	1.04	4	1.06	1.26	1.07	1.03
5	1.00	1.00	1.00	1.00	5	1.00	1.00	1.00	1.00	5	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	6	1.00	1.00	1.00	1.00	6	1.00	1.00	1.00	1.00

Since all of the lineups had 8 members, nominal size = 8, which should serve as an upper bound on estimated effective sizes (but not on estimated functional sizes, see Chapter 6).

Table 8.11 Estimates of lineup size in lineups of varying target-foil similarity and target distinctiveness.

It is clear from Table 8.11 that none of the lineups approaches an effective size anywhere near the notional maximum of 8. This may be a result of the fairly elaborate verbal descriptions given to witnesses - descriptions given in real cases where witness identity is in question are unlikely to be as detailed as this - and the effective sizes given above may consequently be underestimates. A replication, using less detailed descriptions, would test this explanation. Alternatively, it may be that the image set used to generate the lineups is too small to provide foils sufficiently similar in appearance to construct lineups of a reasonable effective size. However, examination of the distribution of witness choices across lineup foils does not support this interpretation: Table 8.12 reports the distribution of choices for the 18 lineups,[313] and shows that some foils attract a disproportionate number of choices. Since foils are 'equally similar' to the target (i.e. have almost identical similarity scores), it cannot be the case that there are too few 'sufficiently similar' foils.

			Lineup member												
No.	Distinct.	Simil.	1	2	3	4	5	6	7	8	Total	χ^2	E	Lower	Upper
1	*High*	1	**10**	0	9	0	2	0	1	0	22	45.64	2.60	3.57	2.05
2	*High*	2	0	1	**16**	9	3	0	0	2	31	59.58	2.74	4.10	2.05
3	*High*	3	1	1	1	**19**	4	1	0	0	26	79.85	1.97	3.38	1.39
4	*High*	4	0	0	0	0	0	**29**	2	0	31	187.06	1.14	1.37	0.97
5	*High*	5	0	0	0	0	0	**28**	0	0	26	182.00	1.00	1.00	1.00
6	*High*	6	0	**30**	0	0	0	0	0	0	30	210.00	1.00	1.00	1.00
7	*Moderate*	1	**11**	0	0	0	8	6	0	0	31	69.39	2.47	3.42	1.93
8	*Moderate*	2	0	1	0	**26**	0	0	0	0	27	173.59	1.08	1.26	0.94
9	*Moderate*	3	0	**15**	0	0	14	1	0	0	30	82.53	2.13	2.44	1.89
10	*Moderate*	4	0	0	0	0	0	0	1	**25**	26	166.62	1.08	1.27	0.94
11	*Moderate*	5	0	0	0	**30**	0	0	0	0	30	210.00	1.00	1.00	1.00
12	*Moderate*	6	0	0	**23**	0	0	0	0	0	23	161.00	1.00	1.00	1.00
13	*Low*	1	1	0	10	1	0	0	0	**18**	30	83.60	2.11	2.85	1.68
14	*Low*	2	0	0	**21**	0	1	0	1	0	23	131.09	1.19	1.57	0.96
15	*Low*	3	0	0	0	**21**	3	0	0	2	26	113.69	1.49	2.19	1.13
16	*Low*	4	0	**30**	1	0	0	0	0	0	31	201.52	1.07	1.22	0.95
17	*Low*	5	0	0	0	**31**	0	0	0	0	31	217.00	1.00	1.00	1.00
18	*Low*	6	0	0	0	0	0	0	0	**26**	26	182.00	1.00	1.00	1.00

Shaded cells indicate lineup positions taken by suspects. Distinct. = distinctiveness; Simil. = similarity. Lower = lower limit of 95% confidence interval around estimate of 'E', Upper = upper limit of same. χ^2 values, and confidence intervals, are calculated on the basis of the theoretical development in Chapter 6. All χ^2 values are significant at p = 0.001

Table 8.12 Frequencies of mock identifications, for each of the 18 lineups in Study 6.

[313] This is a large table, and might appear better suited to an appendix for reproduction. I include it in the text of the chapter, since the aim here is to evaluate measures of lineup fairness, and the method I pursued in Chapter 6 relied on visual examination.

Statistics assessing lineup fairness are also shown in Table 8.12: they suggest that all of the lineups suffer from poor effective size, and this is borne out by visual inspection. The identification frequencies, in turn, suggest that foils do not equally resemble the verbal description: distributions are not uniform, despite the fact that foils were chosen to have equal similarity scores.

The central intention in Study 6 was to investigate the relation between the PC-based similarity measure and measures of lineup fairness. The results clearly establish that there is a strong relation, but it is not clear that similarity scores can be used to construct lineups with high effective sizes. This requires investigation elsewhere.

To complete the report of empirical work undertaken for this thesis, I turn to an examination of the effects of facial similarity in a simulated identification scenario.

Study 7: Facial similarity and identification accuracy

I argued in Chapters 4 and 5 that facial similarity has received very little attention in either the face recognition or witness identification literatures. On occasions where it has been investigated, it has proved to be a variable of substantial import (see Chapter 4, page 90 onwards). We know that it is strongly associated with indices of lineup fairness (Malpass & Devine, 1984; Study 6 in the present chapter), and it appears to affect identification accuracy in simulated identification scenarios. Studies that have investigated the impact of similarity in identification scenarios have typically used indirect measures of similarity (for example, ratings made by independent judges), or *a priori* similarity classifications. In the present study, I investigate the effects of similarity on identification ability with a direct measure of facial similarity, namely PC-based spatial distance.

Although foil-target similarity has rarely been investigated in the identification literature, it is a pivotal topic in theoretical discussion of identification parades. Navon (1990a, 1990b) and Wells and his colleagues (Lindsay & Wells, 1980; Wells, 1993; Wells, Seelau, Rydell, & Luus, 1994), among others, take physical similarity to be a key component in the construction and evaluation of lineups. Researchers in the field differ in their understanding of the way that target-foil similarity is likely to affect identification ability. On the one hand, target-foil similarity is seen as an optimal-function problem: If similarity is too high, suspects will be protected when they are innocent, but guilty suspects will go free. If similarity is too low, guilty suspects will be identified with ease, but so will innocent suspects. The problem is therefore to find a similarity function that maximises identification of guilty suspects, and protection of innocent suspects. According to Wells and colleagues, however, this strategy is bound to fail. There are several reasons that this is the case, but central is the mistaken reliance on the 'match-to-suspect' nature of the strategy. What is required instead is a 'match-to-description' strategy, aimed at ensuring that 'propitious heterogeneity' of parade members is

achieved. Opinion in the psycho-legal literature is divided on this set of claims, and since I discussed the dispute at length in Chapter 4, I will not consider the matter any further here. I simply wish to point out that the 'match to description' approach is not uncontested. This is important in the present study, and indeed to the entire empirical project reported in this thesis, since the measure of facial similarity advocated here assumes the 'match to suspect' strategy.

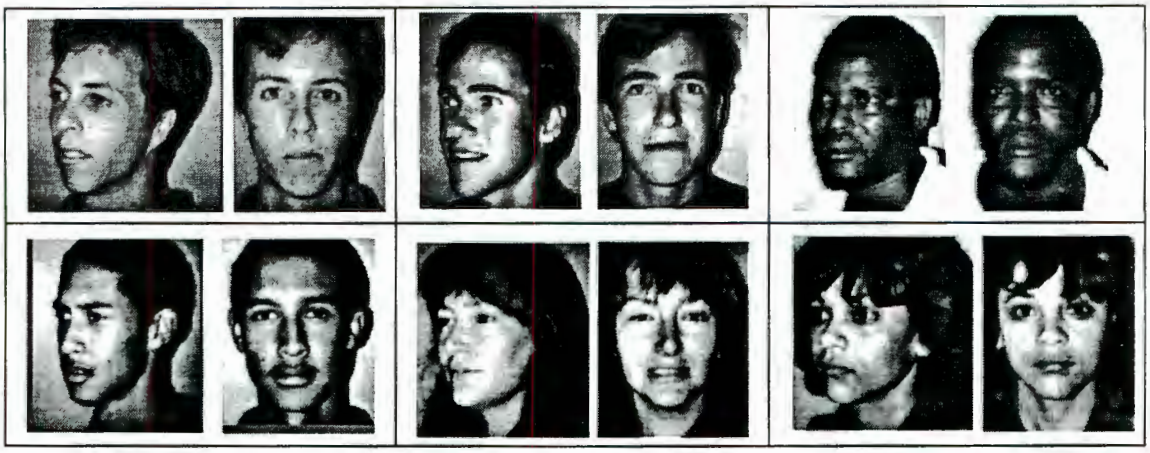
A few important methodological considerations are worth outlining prior to the report of Study 7. Work reported in the identification literature has shown that simulated identification scenarios which use lineups as recognition tests need to bear in mind the distinction between 'target present' and 'target absent' lineups. Both are necessary if one wishes to correctly evaluate identification accuracy, and especially if one wishes to obtain 'diagnosticity' estimates (see Chapter 4). Secondly, one of the most significant contributions of witness identification research has been the development of 'sequential lineups'. In Chapter 2 I argued that this development has had a major impact on police practice in parts of the U.S.A.. Any assessment of the effect of facial similarity on identification ability needs to use both forms of lineup (i.e. simultaneous and sequential). Study 7 incorporates both of the considerations outlined in this paragraph, i.e. it uses target-present and target-absent lineups, and it uses simultaneous and sequential lineups.

Subjects

Subjects were 22 Psychology 3 students, and 46 Psychology 2 students, at the University of Cape Town. Psychology 3 students participated in the experiment during a tutorial, and Psychology 2 students participated during a lecture.

Materials

Two sets of three face images were chosen from the frontal set of 278 images, along with corresponding profile views of these images. A set of face images was presented to each subject at the first stage of the experiment, in document form. The frontal images were also embedded in lineup arrays, which were presented to subjects at the final stage of the experiment. The lineup arrays varied in terms of target-foil similarity, so as to constitute three levels of similarity. This was achieved by selecting foils who fell within 0 - 10, 45 -55 or 90 - 100 percentile points of the target, on the similarity measure. Frontal and profile views for each of the faces are shown as Figure 8.14



Sets are shown in rows.

Figure 8.14 Sets of images used as targets in Study 7

Two documents were prepared to present to subjects at the first stage of the experiment. Each document contained one of the sets of images, along with fictitious descriptions of each of the people represented by the images. Subjects were required to read these descriptions, and having read them, to write three further facts they believed to be probably true of each of the three people. The documents are reproduced in Appendix J. These documents presented information which was later tested with lineup arrays. The point of requiring subjects to read descriptions, and to write three further facts, was to disguise the nature of the experiment: that is, to make it appear like a 'first impressions' study, rather than a simulated identification study.

Simultaneous and sequential lineup arrays were then created for each of the three images: each of these arrays either contained or omitted the relevant face image, thus constituting the 'target-present', 'target-absent' manipulation. Arrays were combined into booklets, three at a time: each target was 'represented' in one of these arrays (either in target-present or target-absent form). In the case of simultaneous parades, these were simply stapled together, along with an instruction page. In the case of sequential parades, a mini-booklet was created for each of the three arrays. One image was printed on each page of the mini-booklet. Three mini-booklets and a page of instructions were inserted into an envelope. Instructions attempted to ensure that sequential lineup arrays were completed as sequential tasks, and those issued for the second set of images are set out below.[314]

> Earlier, you were provided with pictures and descriptions of three students. What I would like you to do now is to point them out - if they appear - in the collections of pictures that follow. Please read the instructions carefully.
>
> 1 Over the page you will find the first collection of numbered pictures, arranged in sequence (I have marked it as 'A'). Please start with the first picture and decide whether it is *Cassiem*. If it is, indicate which number is below his picture. If it is not Cassiem, turn the page. Decide whether this second picture is *Cassiem*. Continue in this way until you have either found *Cassiem*, or until there are no more pictures. *Do not turn back at any stage to look at pictures you have already rejected.* If you have not found *Cassiem* at

[314] Note that points 2 and 3 are shortened for presentation here, since they repeat instructions given in point 1.

the end of the sequence (he may or may not be in the sequence), you must indicate that he is not present. If you do find him in the sequence, write the number which appears below his picture in the space below, and immediately stop the task (i.e. do NOT look at any pictures which appear later in the sequence).

Cassiem does not appear ☐ *Cassiem* is number __

2 Turn the page to the next collection of numbered pictures (I have marked it as 'B'). Please start with ... later in the sequence).

Susan does not appear ☐ *Susan* is number __

3 Turn the page to the next collection of numbered pictures (I have marked it as 'C'). Please start with ... later in the sequence).

Astrid does not appear ☐ *Astrid* is number __

Instructions for simultaneous lineup conditions can be found in Appendix K, along with reproductions of the arrays used in simultaneous and sequential conditions.

Design

Study 7 was a 2x2x2x3 factorial experiment, with one dependent variable. Factors were 1) **Lineup structure** (simultaneous, sequential); 2) **Target presence** (present, absent); 3) **Image set** (a, b); 4) **Target-foil similarity** (high, moderate, low). Lineup structure, Target presence and Image set were between-subjects factors, while Target-foil similarity was a repeated measures factor. Assignment of Image set, Target presence and Lineup structure conditions was random. The design of the study is shown in Figure 8.15, excluding the image set manipulation.[315]

	Present		Absent		Target presence
	Simultaneous	Sequential	Simultaneous	Sequential	Lineup structure
High	12	20	17	14	
Moderate	19	20	15	14	
Low	15	14	19	20	
Similarity					

Cell entries are numbers of subject responses, and not numbers of subject per se, since subjects completed three tasks. (The 'similarity' factor was a repeated measure).

Figure 8.15 Design of Study 7.

Procedure

The experimental procedure differed slightly across the groups of subjects, and this is worth detailing. Psychology 3 students were recruited from groups attending voluntary additional statistics tutorials at the end of the academic year. These groups varied in size, ranging from 2 to 7. At the beginning of the tutorial they were given a document containing frontal and profile views of targets, and were allowed five minutes to complete the task contained in the document. There were two documents

[315] Two image sets were used in order to check variation of results across images. Effects were not specific to image sets, and this manipulation will not be discussed further.

designed for this stage of the experiment (as detailed in Materials), and subjects were randomly assigned one of the documents. A statistics lesson then ensued, and lasted 30 minutes. At the end of the lesson, subjects were asked to complete the second part of the experiment (they did not know that there was a second part), and were randomly assigned an array booklet, which contained either simultaneous or sequential lineup array tasks, corresponding to the image set they had received at the beginning of the experiment. Some of the arrays in these tasks contained the target, and others didn't. The assignment of target present and target absent arrays had been effected randomly, in the development of the experimental materials.

Psychology 2 students were addressed at the beginning of a year-end course lecture, and asked to participate in a study of face recognition and perception. Each subject was then handed two envelopes. One envelope was marked

> Open this envelope when you receive it. Complete the task inside it, place the completed task back in the envelope, and seal it.

This envelope contained the first stage of the experiment, namely the document discussed in Materials. The second envelope was marked 'Do not open this envelope until instructed to do so', and was sealed. Subjects completed the first task, after which the lecture commenced. The lecture lasted 30 minutes, after which subjects were instructed to open the second envelope. They then completed the lineup array tasks, which were contained in the envelope. Lineup structure, Target presence and Image set conditions had been randomly distributed across envelopes, so assignment of subjects to conditions was also random.

Results

Similarity, lineup structure, and target presence all proved to have significant effects on subject performance in lineup tasks. Subjects made better decisions with sequential lineups, but only when targets were absent; and low-similarity lineups improved accuracy of identifications in sequential and simultaneous lineups, in both target-absent and target-present conditions. Image set had no effect on subject performance, and since this manipulation was intended to check the generalization of findings across images, it will not be discussed any further in this chapter.

Results from lineup tasks may be assessed in terms of correct identification decisions (i.e. to treat identifications of the perpetrator when he is present as equivalent to witness indications that the perpetrator is not present when he is indeed not present), but it is generally more useful to classify results in relation to target presence/absence. Results are typically reported in the terms set out in Table 8.13

	Target status	
	Target present	Target absent
Identifies suspect	hit	n/a
Identifies foil	incorrect id.	incorrect id.
Identifies no-one	incorrect rejection	correct rejection
Witness identification decision		

Note: in some instances, researchers classify 'identifies suspect' under Target absent as an incorrect identification, or 'false alarm'. This happens when one of the foils is pre-designated as the 'suspect' - see the discussion below.

Table 8.13 Identification decisions and their outcomes in simulated identification scenarios.

Results are thus better considered separately, at least initially, for target present and target absent lineups. Table 8.14 reports results for Study 7, according to lineup structure, and target-foil similarity. χ^2 tests of association between lineup structure and identification performance are also reported.

	Target present							Target absent						
	High similarity			Moderate similarity			Low similarity			High similarity			Moderate similarity			Low similarity		
	Seq	Sim		Seq	Sim		Seq	Sim		Seq	Sim		Seq	Sim		Seq	Sim
hit	10	5	hit	8	5	hit	10	13									
incorrect id	6	3	incorrect id	1	4	incorrect id	2	0	incorrect id	6	15	incorrect id	7	6	incorrect id	3	8
incorrect rej	4	4	incorrect rej	11	9	incorrect rej	2	2	correct rej	10	4	correct rej	7	9	correct rej	17	11
$\chi^2 = 0.71$; df = 2; $p > .70$			$\chi^2 = 2.52$; df = 2; $p > .28$			$\chi^2 = 2.40$; df = 2; $p > .30$			$\chi^2 = 7.04$; df = 1; $p < .01$			$\chi^2 = 0.3$; df = 1; $p > .50$			$\chi^2 = 3.54$; df = 1; $p < .06$		

Seq = sequential; Sim = simultaneous; id = identification; rej = rejection.

Table 8.14 Identification decisions in the 12 lineups used in Study 7

Table 8.14 shows that there is little difference in identification accuracy between simultaneous and sequential lineup structures when targets are present, but there is a difference when targets are absent. In particular, subjects completing simultaneous lineup tasks are apt to make more incorrect identifications (false alarms) when targets are absent, than when targets are present. This trend appears to be the case for lineups of high and low target-foil similarity, but not for lineups of moderate target-foil similarity. In addition, there appears to be an effect for facial similarity across lineup conditions - from visual scrutiny, low similarity appears to be associated with a great many hits, and few mistakes, in both sequential and simultaneous conditions. This impression is investigated more rigorously below.

One of the most useful ways of thinking about lineup identifications is in terms of Wells and colleagues' 'diagnosticity ratio' (see Chapter 6). This ratio reflects the likelihood that a guilty suspect

will be chosen, when present in a lineup, over an innocent suspect being chosen, when the perpetrator is not present. Although this ratio makes perfect sense when one has police lineups in mind (the police always have a suspect, and this suspect is either innocent or guilty), it does not apply as easily to lineups used in simulated identifications. This is because simulated identification scenarios always have a 'real' perpetrator, but never have a 'real' suspect, in the sense that police have a 'real' suspect. In order to calculate the diagnosticity ratio in a simulated identification scenario, researchers usually designate one of the lineup members in target-absent arrays as the 'suspect'. This method seems a little opportunistic to me,[316] and I decided to approach matters a different way. Instead of treating one of the members of the target-absent lineup as the suspect, I treated all identifications of foils as informative, and totalled the number of identifications in these lineups (totals are shown in Table 8.14). In a real lineup, the identifications accruing to only one of the lineup members would be considered (since there is only one suspect[317]), so I divided the total number of identifications in target-absent lineups by the number of lineup members. This estimate is used in the calculation of diagnosticity, in place of the number of identifications attracted by the 'designated suspect'. Diagnosticity ratios are presented in Table 8.15, for sequential and simultaneous lineups, across the three similarity conditions.

	Similarity		
Lineup structure	High	Moderate	Low
Simultaneous	4.2	5.5	16.5
Sequential	10.7	6.4	38.1
Total	**6.3**	**6.0**	**22.1**

Values in the row marked 'total' are calculated over both conditions.

Table 8.15 Diagnosticity ratios for lineups constructed to have varying target-foil similarity.

It is clear from the table that diagnosticity varies across similarity conditions, and across lineup structure. It also appears that lineup structure and similarity factors interact, so that differences between simultaneous and sequential lineups are greatest when similarity is low, and smallest when similarity is moderate. These observations should be subjected to examination by inferential statistics. It is possible to compare simultaneous and sequential lineups across similarity conditions, using methods I developed in Chapter 6, since these comparisons are of independent diagnosticity ratios. However, the methods developed there are suitable for large samples, as they rely (in part) on

[316] See the discussion of Navon's arguments, starting on page 75 of Chapter 4.

[317] I assume that there is one suspect, although police do hold multiple-suspect lineups. See Chapter 3.

approximation to the normal probability distribution. Table 8.16 reports tests of significance, as well as confidence intervals for diagnosticity ratios.

Similarity	Significance tests					95% Confidence intervals					
	Sim. d	Seq. d	χ^2	df	prob.	Sim. d'	Upper limit	Lower limit	Seq. d'	Upper limit	Lower limit
High	4.2	10.7	0.45	1	p > 0.5	3.6	8.37	1.56	6.8	11.8	3.9
Moderate	5.6	6.4	0.007	1	p > 0.5	3.7	7.62	2.51	4.4	7.6	2.5
Low	16.5	38.1	0.19	1	p > 0.5	11.3	36.6	3.5	17.0	40.7	7.1

All statistics reported in this table are as described in Chapter 6

Table 8.16 Tests of significance, and confidence intervals, for diagnosticity ratios calculated on simultaneous and sequential lineups.

Values reported in the table suggest that differences between diagnosticity ratios are not statistically significant. This can be seen from the χ^2 tests of differences between d, and from the overlapping confidence intervals for d'. These results may serve merely to underscore that the methods developed in Chapter 6 are appropriate for large samples: for small samples, statistical power will be very low.[318] Certainly, χ^2 tests of association reported in Table 8.15 show that simultaneous and sequential lineups exhibit different proportions of identification error.

The design of Study 7 demands an investigation of similarity effects across target presence and lineup structure manipulations. There are two problems in implementing such an analysis, though. The categories that the dependent variable takes differ across target-present and target-absent lineups, making comparison difficult. This can be overcome by reclassifying identification decisions as 'correct' or 'incorrect' (which loses information, but achieves comparability), and in the analysis reported below, this variable is called 'identification accuracy'. In the second place, the design of Study 7 incorporates a repeated-measures factor (similarity), which, like all other variables in the design, is categorical. Log-linear analysis is the analytic method of choice for exploring main and interaction effects of categorical independent variables on categorical dependent variables, but there is no general method for designs which use repeated measures.[319] I proceeded to analyse the data for

[318] This can be seen from the formula for the standard error of ln(d'), below: As cell sizes increase, so the standard error decreases, and, by implication, power increases.

$$S[\ln(d')] = \sqrt{\left(\frac{1}{n_{11}+0.5} - \frac{1}{n_{\cdot 1}+0.5} + \frac{1}{n_{21}+0.5} - \frac{1}{n_{\cdot 2}+0.5}\right)}$$

[319] There are 'table folding' methods, which fit *symmetry*, *quasi-symmetry*, and *quasi-independence* models (see Agresti. 1990; and Kennedy, 1992), but these appear to be appropriate for one-factor repeated designs only. In the present design, there are two independent factors in addition to the repeated factor, and there does not appear to be a widely recognised technique for dealing with such a design.

Study 7 with standard log-linear techniques (i.e. assuming all factors to be independent): this appeared to be the only option which would provide a full evaluation of the study,[320] even though it does not take the dependency in the data into account.

Tests of all k-factor interactions suggested that 2 and 4 factor interactions should be included in the model (2 factor interactions: $\chi^2 = 23.7$; df = 9; $p < 0.006$;[321] 4 factor interaction: $\chi^2 = 4.65$; df = 2; $p < 0.1$), but tests of partial association suggested only a two way interaction between similarity and identification accuracy ($\chi^2 = 17.54$; df = 9; $p < 0.001$). Specific models tested against each other revealed that the model {Similarity X Identification accuracy}[322] provided an adequate fit, and was simpler than any rival models. Table 8.17 summarises the model identification and evaluation.

No.	Model spec.	χ^2	df	prob.
1	I, St, P, Si	37.20	18	p > 0.005
2	I x Si	19.92	18	p > 0.33
3	St x P, St x Si, I x St, I x P, P x Si	31.05	11	p > 0.001
4	I x Si, St x P, St x Si, I x St, I x P, P x Si	13.51	9	p > 0.14
5	St x I x Si	14.04	12	p > 0.29
6	St x Si x I, St x P x I, St x I x Si, I x P x Si	4.66	2	p > 0.09

I = Identification accuracy; Si = Similarity; St. = Lineup structure; P = Presence of target.

Table 8.17 Key log-linear models tested in the log-linear analysis of Study 7

Model 4 is clearly preferable over model 1, since it fits the data better, and the incremental change in χ^2 is significant. This implies that at least one 2-factor interaction will be required to fit the data. Model 2 is preferable over model 3, by the same criterion of incremental change in χ^2. A comparison of models 2, 3 and 4 shows that it is necessary to include the interaction between Similarity and Identification accuracy, but no other 2-factor interactions. (This is corroborated by the test of partial associations, reported above). It is not easy to choose between Models 2, 5 and 6 on statistical criteria, but the parsimony of Model 2 makes it the most attractive.

The conclusion from the log-linear analysis is that the interaction between Similarity and Identification accuracy is a sufficient basis on which to understand the results of Study 7. This interaction is shown in terms of cell frequencies, in Table 8.18 below:

[320] The analysis must therefore be viewed with some skepticism. The central assumption broken here is that of statistical independence. One of the most likely consequences of breaking this assumption is that statistical tests of significance used in identification and testing of the log-linear model will be conservative, i.e. the model will be over-simplified.

[321] All χ^2 values reported in this section are estimates produced by maximum likelihood estimators.

[322] This is a hierarchical model, and interactions thus assume the presence of all earlier interaction and main effects.

Similarity	Identification decision: Incorrect	Correct
High	38 (57%)	29 (43%)
Moderate	40 (56%)	32 (44%)
Low	19 (26%)	53 (74%)

Percentages are calculated row-wise.

Table 8.18 Frequencies for the interaction between facial similarity and identification accuracy.

The interaction is explicable almost entirely in the deviation of the low similarity condition from other conditions: in this condition, correct identifications were much more frequent than in high and moderate similarity conditions. A re-examination of Table 8.14 shows that correct identifications in low similarity lineups are more frequent in both target-present and target-absent conditions.

Discussion of Studies 6 and 7

The major aim of empirical work reported in this thesis has been to develop a direct measure of facial similarity. I showed in Studies 2 - 5 that a measure based on a principal component analysis of face images succeeds in an important respect, namely that it corresponds reasonably well to human judgements of facial similarity. Since the justification for developing a measure of similarity was derived from analysis of a legal problem - how to assess the fairness of identification parades - this was not sufficient evidence in favour of the measure. I needed, in particular, to show that the measure of facial similarity would also correspond to measures of lineup fairness developed in the psycho-legal literature. Study 6 investigated the measure in this respect, using the well-tried 'mock witness' technique, and several measures of lineup fairness derived from this technique. Results showed that the measure of facial similarity is strongly related to measures of lineup fairness. The results were complicated to some degree by slightly anomalous results from a 'facial distinctiveness' manipulation. The complication is not too serious: there was a near-monotonically decreasing relation in all distinctiveness conditions between similarity and lineup fairness. In addition, the PC-based measure of 'facial distinctiveness' has proved very difficult to understand in several of the studies reported in this chapter, particularly since it appears unrelated to subject judgements of distinctiveness.

At any rate, the measure of facial similarity is strongly related to measures of lineup fairness, and may prove to be a useful proxy for these more expensive and less direct measures. Ideally, one should also be able to use the measure in foil selection, particularly for photo-lineups. The measure would

(theoretically) allow one to structure the similarity of a lineup in an 'objective' manner. Results from Study 6 do not support such a use for the measure (but later investigations may, of course, find differently). The distribution of witness identifications across foils was decidedly non-uniform, even in high similarity conditions, and even though foils in these conditions had near-identical similarity scores in relation to the target. This may have been due to the highly individuating descriptions subjects were given of targets, but it may also show the failure of the 'match-to-suspect' strategy that the similarity measure adopts. It would be interesting to compare a 'match-to-description' foil selection strategy directly with a method based on the facial similarity measure. In general, lineups in Study 6 exhibited low levels of fairness, and later work will need to investigate the similarity measure with a wider range of fairness scores.

Study 7 did not directly examine the utility or validity of the similarity measure. Instead, the measure was used there to investigate the effect of similarity on witness identifications, using a simulated identification scenario. Results from the study showed that similarity is strongly related to witness accuracy, but in an unanticipated direction.[323] Low similarity lineups led to greater accuracy, in terms of hits and correct rejections, than moderate or high similarity lineups. This finding bears out Wells' recent contention (Wells et al., 1994) that high similarity lineups may not provide witnesses with 'propitious heterogeneity'. The finding also lies quite uneasily next to the finding from Study 6 that high similarity lineups are associated with greater lineup fairness. It may be that the conventional way in which 'lineup fairness' is understood and investigated in the psycho-legal literature is mistaken. Fairness is usually assessed with mock witness tasks, and is measured as bias towards (or against) the suspect, or in terms of the number of 'plausible' foils in the lineup. The mock witness task assumes that a lineup is unfair if a witness is able to identify the suspect with only a brief description of the suspect to hand. What constitutes 'brief' has not been investigated, and this may be overdue: one expects a witness who has a detailed description to succeed, so we need to know more about the relation of the description to mock identification accuracy.

Study 7 also corroborated findings from previous studies in the field regarding the utility of sequential lineups. Sequential lineups were associated in this study with fewer false alarms than simultaneous lineups, while securing the same number of 'hits'.

In sum, the key findings from Studies 6 and 7 are that the PC-based measure of facial similarity may be able to 'stand-in' as a proxy for standard measures of lineup fairness, but that these standard

[323] It is not correct to claim that this result was totally unanticipated. Wagenaar & Veefkind (1992) reported results which showed that greater similarity led to more false alarms, and decreased recognition accuracy. Much legal discussion, however, has assumed that greater similarity will lead to greater fairness.

measures may themselves need re-examination in the light of the relation between lineup similarity and identification accuracy uncovered in Study 7.

Chapter 9 Conclusion

In this, the final chapter of the thesis, I have several aims. In the first place, I wish to review the cardinal arguments of earlier chapters, and to trace their evolution in the thesis. It is clear from the *a priori* and empirical investigations reported in Chapters 6, 7 and 8 that certain arguments and claims need to be reshaped, or moderated. Secondly, this chapter is traditionally set aside for summary and evaluation of findings, and I will not disappoint in this respect: In particular, I will attempt to provide a summative account of research on the proposed measure of facial similarity, including its potential use in the assessment of parade fairness. Finally, I will return to one of the key meta-theoretical engagements of the earliest part of the thesis, which is the notion of an 'applied psychology': my brief will be to speculate about the practical use of the similarity measure in several aspects of police and legal work.

I argued at the beginning of the thesis that research on identification evidence and identification parades is a kind of 'applied psychology'. We must be careful, though to understand the nature of 'applied research': It is all too often interpreted as the 'application of basic science to real world matters', or as some other near-tautological pursuit. I argued earlier that this is frequently mere justificatory rhetoric on the part of researchers. Most of what passes as 'applied psychology' is never applied, and would be better labelled as 'applicable psychology'. The label is adopted on the basis of the distal origin of the research problem, and because recent cognitive psychology has greatly favoured research which attempts to shed itself of the laboratory. A psychology which is applied, on the other hand, will take seriously the notion that 'application' is not inherent to research; it needs to be actively pursued.

A useful point of departure, in the case of applied witness research, is to take bearings on the distinction made by Wells (1978) between 'system' and 'estimator' variables. System variables are those under the control of the criminal justice system, and research on these works close to the form (eventual) application will take: the trick is to find something about which something can be done. Choosing a system variable as the focus of the research does not guarantee application, though: in Chapter 2 I reviewed a discussion by Wells (1986), which outlines potential strategies for promoting the application of witness research, and I think I made clear in that review that there are many obstacles in the way even of research which carefully targets 'system variables'. Nevertheless, the choice of system variables as objects for research investigation appears to be a useful beginning strategy. The research reported in this thesis utilised this strategy.

The present work sets itself up as applied research, and its amibitions in this respect should consequently be subjected to inspection.[324] Certain questions can be posed in respect of the 'applied' orientation of empirical research reported in this thesis. The present research stopped well short of 'application': limitations are imposed in the first place by the inherent academic nature of the thesis, and the ideas pursued here are, in the second place, at an early stage of development. Nevertheless, I believe that results from the empirical investigation of the similarity measure show it to be of potential use in several aspects of police and legal practice. I will outline a few of these potential uses at the end of the chapter.

The cardinal argument regarding the understanding of 'applied psychology', at least for purposes of continuity of the thesis, was that researchers need to proceed carefully at the stage of problem definition. The justification for applied research is typically a practical problem which originates outside of the discipline in question, and an important part of an applied project must be to understand the problem in terms of the discipline or context being entered. There is no point in answering a problem which is phrased incorrectly. This conclusion led me to a review of South African law on identification parades, as well as a brief look at police practice.

The review of South African law was approached in the spirit of a kind of legal positivism. That is, the review was an attempt to construct a coherent, rule-based interpretation of the law. This was achieved through the inspection of manifest law, i.e. statutes, case law and police documents. There was no attempt to assess legal practice in a more grounded manner. Legal practice with respect to identification parades may turn out to differ greatly from this picture: we may find that overt characteristics of judges and defendants partly or largely determine the conduct and evaluation of parades (as they certainly have in other aspects of legal practice in South Africa - see Murray, Sloth-Nielsen & Tredoux, 1989). As far as the review is concerned, it cannot be claimed that a completely coherent picture was obtained. There are many contradictions, and points of disagreement. It is unclear where many regulations originated, and it is clear that different interpretations are possible. This, of course, is not unusual in a complex knowledge system like case law.

The review of South African law on identification parades showed that the chief concern of the courts has been to regulate procedure. Courts have gone to great lengths to ensure that accused persons are protected against suggestive parade arrangements, and to rule out the possibility that witnesses make identifications by virtue of having the suspect 'pointed out', however inadvertently this might occur. Do example, the officer charged with conducting the parade should not even know the identity of the suspect, in case he accidentally communicates this to the witness. Many other strictures are in place, and were discussed at some length in Chapter 3.

---

[324] Lest Byron's adage ring true: 'In every trade a man must do his apprenticeship, save censure'.

There has been little concern, by contrast, with the structure of the parade. The question of how the parade should be assembled is not even addressed, even though the review of comparative legal and psychological literatures in this thesis showed that there are many alternatives to the traditional 'lineup' (e.g. the sequential lineup pioneered by Lindsay & Wells (1985), and the casual 'parade' of people in a room, as used in some Scandinavian countries). In addition, few legal authorities have thought carefully about the nature of the task embodied in the identification parade: the central unanswered question in this respect appears to be whether it should be treated as a test of reliability or as an evidentiary procedure in its own right, i.e. one that gathers independent information that points to the innocence or guilt of the suspect. This question has been addressed in the psychological literature, and there are some interesting arguments, but there is a clear lack of consensus.

The most important aspect of parade structure, however, concerns the physical constitution of the parade. The courts have specified that the members forming the parade should be of same height, build and physical appearance as the suspect, but no advice is given concerning the evaluation of this requirement. In police practice, the officer charged with conducting the parade is simply required to certify that this is the case: he is not required to provide evidence that will point to this conclusion. He may arrange for a photograph to be taken, but is not required to do so.

I identified this lacuna in the legal treatment of identification parades as a major problem, but wish to add at this stage that the problem runs deeper than simply assessing the similarity of parade members. The results of Studies 6 and 7 showed that the relation between similarity, lineup fairness and identification performance is certainly not self-evident, and may well be highly complex. Even if courts were able to assess similarity, they would be faced with the further question of deciding what degree of similarity is required. This kind of question is not amenable to investigation by the courts - or at least not under its present epistemological disposition. It is a question more amenable to investigation by a discipline like research psychology.

The overriding aim of providing the review of South African law was to better understand the problems that identification evidence pose, in context. I argued that one of the key difficulties appears to be the inability to assess the physical similarity of parade members, and I took this observation as partial motivation for the empirical work of the thesis.

Other motivations derived from the review of the psychological literature on identification parades. Psychological work on identification parades has been very productive, if one considers the small number of researchers actively involved in the field, and the relatively small corpus of publications. I argued earlier that measures of lineup fairness and lineup size represent some of the most useful work in this regard, but that the work is flawed by a neglect of the probability assumptions underlying the measures. Chapter 6 constituted an attempt to bolster the measures by tying them to their basis in probability theory. Specific recommendations were made regarding ways to use statistical inferential

reasoning in the interpretation of the measures. This included developing a new measure of lineup size, 'E': earlier definitions of lineup size proved intractable to treatment by inferential statistical techniques, and 'E' was developed in order to be tractable to inferential techniques. I also investigated the interpretation of Bayesian measures known as 'diagnosticity' and 'information gain', and made recommendations regarding the application of inferential statistical thinking to these measures.

Many of the recommendations made in respect of the application of techniques of statistical inference to lineup measures were tempered by the stricture that they be used with fairly large samples. Empirical work in Studies 6 and 7, which used fairly small samples, showed the dependence of the measures on large sample sizes. This stricture, and its confirmation, may make inferential techniques a less attractive option to identification research, since much of the research is based on (relatively) small samples.[325]

Studies 6 and 7 raised additional questions about the measures of lineup fairness and size. Study 6 showed that similarity covaries with both fairness and size: lineups of greater target-foil similarity had greater functional and effective sizes, and led to fewer correct mock identifications of suspects. Study 7, on the other hand, showed that greater target-foil similarity led to lower accurate identification rates, in both suspect-present and suspect-absent lineups. This constitutes something of a conundrum for research on identification parades: fair(er) lineups lead to lower identification accuracy! The conundrum may derive from what is an unexplored aspect of the mock witness task (this task underlies all of the measures of lineup fairness and lineup size).

The mock witness task determines the likelihood that a witness, armed only with a brief written description, is able to guess the identity of the suspect from visual inspection of the parade. The role that the written description plays in the performance of mock witnesses is almost completely unexplored in the literature. Clearly, a more detailed description will lead to a greater likelihood of accurate identification than a less detailed description, and by implication also to lower estimates of lineup size and lineup fairness. This makes the measures of lineup fairness and lineup size arbitrarily dependent on the description. An implication of this is that the more information a witness has, the less likely it is that a 'fair' lineup can be constructed, in the sense that a lineup is adjudged fair by existing empirical measures of fairness. There is no natural limit on the detail of the description - one cannot talk reasonably of an 'average' witness description, since descriptions may vary in as many ways as there are different opportunities for observation. The situation where the fairness of a lineup is investigated in a *post hoc* fashion - for example, where a lineup has been conducted by police, and

[325] It is, of course, always difficult to know what size of sample qualifies as 'large', or 'small'. In the case of identification research, though, many of the measures are proportions (e.g. proportion of correct identifications), and where these approach 0 or 1, very large sample sizes will be needed to provide accurate estimates of population parameters.

the description originally given to the police is available - could perhaps be said to produce conditions that naturally limit the description, but even this type of situation is subject to complications. For one, descriptions provided by witnesses are rarely likely to be complete records of what is available to them in recall. Certainly, some searching questions need to be asked about the measures of lineup fairness and size, and about the mock witness technique in particular.

The principal empirical project undertaken and reported in the thesis developed and tested a measure of facial similarity. The need for such a measure was argued in the first place on the basis of the review of identification law, but to that could be added the virtual absence of a measure in the face recognition literature. When facial similarity is measured, it is typically through an indirect route, and an *ad hoc* solution to a research exigency.

The theory behind the measure proposed in this thesis is Valentine's multidimensional model for the perceptual representation of faces. The multidimensional nature of the model means that faces are represented in terms of the same underlying dimensions, which in turn sustains well known mathematical measures. These allow - among other things - the calculation of distances between faces in the space, and distances of faces from the origin. Unfortunately, Valentine's model is offered at a conceptual level only, and the dimensions constituting the space are left completely unspecified.

I searched for a suitable solution to the problem of dimensions for the multidimensional model, and evaluated possible candidates in Chapter 5. The search for suitable 'facial' dimensions led to the image processing work of Sirovich & Kirby (1987), and O'Toole and colleagues (e.g. O'Toole & Thompson, 1993). This work pioneered the use of principal component analysis (PCA) with digital images of faces, and appears to solve not only the problem of suitable facial dimensions, but also provides an implementation of Valentine's multidimensional model, in the form of the principal component analysis itself. The derivation of spatial distances from principal component solutions is seamless, and a natural consequence of the mathematical basis of the analytic technique.

The major theoretical proposition of the thesis, then, was that a measure of the facial similarity of any two faces in a particular image set can be defined as the Euclidean distance between these faces in a principal component space. A similar proposition was made in the case of facial distinctiveness. The empirical work set out, by and large, to investigate these propositions. Before I turn to a review of this empirical work, it is useful to note certain limitations pertinent to the underlying theoretical basis of the measures, and some uncertainties about the practical implementation of the methodology.

Of the many distance and other potential similarity metrics available in multidimensional space models (e.g. the Minkowski metric, city-distance spatial measures, dot products), I chose only to investigate Euclidean distance. I leave the utility and theoretical adequacy of other measures as an open question to later research, or other researchers. The technique used in implementing the PCA of

face images appears to break some fundamental mathematical axioms. Faces are standardized with respect only to pupil location, and will differ in many other ways after standardization: combinations of faces treated in this way will not again be 'faces', and this breaks one of the axioms of linear systems. I discussed a particular solution to this problem in Chapters 5 and 8, but argued there that this solution cannot meet the axioms in question, if they are applied strictly. In the end, it is not clear how the axioms should be interpreted. I have settled for an interpretation which requires combinations of face images to be 'facelike', rather than faces themselves.

There are technological limitations, presently, in the practical implementation of the methodology required to derive facial similarity scores from a principal component analysis. All images in the studies reported in this thesis were standardized individually, using image processing software, and this was a lengthy business. After standardization, images were subjected to PCA, and component coefficients captured for each face. Finally, Euclidean distances were calculated between faces. This process is cumbersome, at least in its present form, and solutions to some practical limitations are needed. The standardization procedure is the most time consuming stage in the process, and software capable of automating this process would be very useful. This could be achieved simply enough if pupils could, in turn, be automatically determined for each face. Recent publications in the engineering literature suggest that this will soon be possible (Cheng, Liu & Yang, 1993). The solutions to other implementation problems appear to be uncomplicated, and probably depend only on the availability of resources.

I have repeatedly claimed that the present research - and, indeed, most identification research - is a kind of applied psychology, but this does not mean that theory is unimportant. I have determined the value of the PCA similarity measure according to its correspondence with human perceptions of similarity, and its relation to lineup indices, but it could be said that such an approach is mere pragmatism. What is the relation between the technique used to derive the similarity scores, and the human visual-cognitive system? Surely it cannot be the case that human perceptions of similarity are determined in a manner which is anything like a principal component analysis, or again, like the determination of a Euclidean distance in a multidimensional space? A better approach in the long run, it may be argued, is to develop a theory at the level of cognitive process, and to develop measures from this theoretical groundwork.

There are several possible rejoinders to this position. Although it may seem a bit far-fetched to assert that the PCA technique underlying the similarity measure is a good model of visual-cognitive processes, some support for this proposition can be adduced from the demonstration that the technique can be translated into a (mathematically equivalent) auto-associative neural network (see footnote 249 in Chapter 5). Neural networks are now widely used to model cognitive processes, and their ability to produce 'learning behaviour' of the type that human face recognition behaviour shows,

suggests that there is more to the PCA approach than sheer expedience.[326] There are some, though, who feel that neural network models are in themselves highly dubious accounts of human cognition (see Bechtel, 1991). Perhaps the truth of the matter is that most cognitive modelling is difficult to justify by reference to 'real psychological processes': we rarely have evidence that makes us capable of arguing this point with any degree of competence.

I concede then, that the cognitive-theoretical basis of the PCA approach is a little inchoate, and that I have done little to render it differently. This can be the task for a later, more extensive project.

The empirical tests undertaken in respect of the facial similarity and distinctiveness measures can be separated into two categories. In the first, I investigated the measures directly - that is to say, I examined the relation of the measures to human judgements of facial similarity and distinctiveness, on the same sets of stimulus materials. Several distinct task operationalizations were utilised. Direct investigations of the measures also included an assessment of concurrent validity, in respect of the ability of the measures to accurately discriminate *a priori* similarity classifications of sets of face images. In the second category of test, I examined the relation of the measure of facial similarity to indices of lineup fairness and lineup size. This category of test does not directly examine the validity of the measure: it assumes instead that target-foil similarity determines lineup fairness, *ceteris paribus*, and hence treats the facial similarity measure as a proxy index of lineup fairness. The first five empirical studies fall into the first category, and the final two into the second.

The first two studies produced anomalous results when face ranking tasks were used (i.e. there was correspondence between subject rankings and rankings based on the PC similarity measure, in some, but certainly not all tasks). I argued that the anomalous nature of these results is explicable, on several grounds. I will not detail these again here, except to note that these studies used images (and similarity measures) derived from a relatively small, homogenous base set. Although results from the ranking tasks were unclear, a face pairing task used in Study 2 showed a strong and consistent correspondence between subject judgements of similarity and the PC measure. In addition, a cluster analysis produced a classification which grouped similar faces, and distinguished dissimilar faces, and a discriminant analysis appeared capable of producing highly accurate classifications of race, sex and age groupings.

Study 3 set out to correct significant methodological uncertainties which may have hampered Studies 1 and 2. The most important of these included the collection of a new, larger, and more heterogeneous base set of facial images, and the re-structuring of the similarity judgement task used in Study 2. Similarity ratings were gathered in 11 conditions, using the restructured similarity rating

[326] O'Toole et al. (1994), for example, have shown that the neural network implementation is capable of modelling the 'other-race' effect in face recognition behaviour.

task, and the correspondence between these ratings and the PC similarity measure was generally strong, and always in the expected direction.

Study 4 repeated the investigation of the correspondence between subject ratings and the PC measure, but extended the scope of the inquiry to include different viewing perspectives. Results again generally showed a strong relation between subject judgements and the PC measure. This was true for ratings of frontal faces and for ratings of profile faces. In addition, PC similarity scores based on frontal views were strongly related to PC scores based on profile views of the same subjects (indeed, perhaps more strongly than subject ratings of similarity across viewing perspectives). This is an important test of the measure: frontal views of faces should not be mistaken for the faces themselves, and a similarity measure based only on frontal views runs the risk of making such a mistake - a measure based on frontal views may simply be a measure of picture or image similarity, and not a measure of the similarity of the faces.

In Studies 1-4, judgements of facial similarity exhibited a great deal of inter-subject variation. This is a significant finding in its own right: although several published studies have elicited similarity ratings, only one of these refers at all to this phenomenon, and then only as an aside (Lindsay, 1994). I considered several possible explanations for the high degree of inter-subject variation in Chapter 7, and devised Study 5 as a test of the most troubling explanation. This explanation posits that similarity ratings are inherently unstable, and will vary greatly over rating occasions, even when these are made by the same subject. Study 5 examined the fluctuation of ratings over a three week period, using a simple test-retest design. Findings indicated satisfactory reliability over time, but there were several weaknesses in this design that could be corrected in further investigations of the 'variation phenomenon'. In addition, Study 5 tested only one explanation of the phenomenon; a series of studies aimed at explicating inter-subject variation would be a welcome addition to the literature.

In sum, Studies 1-5 show that the PC similarity measure has considerable promise. The measure is strongly predictive (in general) of human judgements of facial similarity, and this is true over a variety of conditions, sequences, task operationalizations, and viewing perspectives. The measure is also capable of discriminating groupings which, on *a priori* grounds, reflects similarity differences between subjects.

Investigations of the PC distinctiveness measure, on the other hand, produced less promising results. Studies 1 and 2 showed little evidence that this measure corresponds to human perceptions and judgements of facial distinctiveness. Stepwise regression procedures using vectors of component coefficients as predictors, however, appeared able to model human data with some success. Further investigations of distinctiveness, to wit in Study 3, showed a similar lack of correspondence between subject judgements and the PC measure, but stepwise regression analyses appeared to show again that human data can be successfully modelled. I am skeptical of the regression results, though, for reasons

set out at greater length in Chapter 8, and prefer to conclude that facial distinctiveness information is not captured by the PC distinctiveness measure, at least as defined in this work. Alternate definitions of distinctiveness may be the first step in rethinking the PC measure. The distinctiveness of a face is defined in Valentine's multidimensional model as the distance of the face from the origin of the space. PCA effectively rules out an operationalization of this definition, since it determines a solution subject to the constraint that members of the component space are orthonormal (i.e. they are orthogonal, and of unit distance from the origin). Faces are represented equidistantly from the origin, and this means that it is impossible to implement the theoretical definition of distinctivness. There may be ways of transforming component solutions to achieve an operationalization of Valentine's conceptualization of distinctiveness; I leave this possibility to later research, or other researchers.

The two final empirical studies undertaken for the thesis investigated the application of the similarity measure to identification parades. They thus fall into the second of the categories delineated above.

Study 6 explored the relation between the PC similarity measure and the indices of lineup fairness and size discussed at some length in Chapters 4 and 6. A mock witness task was designed, and eighteen photo lineups were constructed in order to examine the relation between similarity and distinctiveness manipulations, and parade indices. Results revealed a definite and strong relation between similarity and lineup indices; this relation was present for each of the measures of lineup fairness and size, and appeared to be monotonic, but the presence of ceiling effects must make conclusions about the nature of the relationship somewhat tentative. Greater similarity led to greater fairness, and to higher estimates of lineup size. It should be mentioned, though, that none of the lineups used in the study produced size estimates anywhere near the nominal number of lineup members, and few, if any, of the lineups produced suspect identification rates that would lead one to believe that they were fair. This was probably due to the detailed nature of the written descriptions mock witnesses used to determine the identity of the suspect. As I argued earlier in this chapter, we need to investigate the significance of these descriptions to measures of lineup fairness and size at much greater length.

One or two remarks about the application of the similarity measure to identification parades are in order here. In the first place, it should be recognised that the 'fairness' of a parade will depend on many things. Procedural irregularities may produce an identification which is palpably unfair (see the newspaper report reproduced on page 58, Chapter 3) even though the structure of the parade is satisfactory in every respect. Parade structure may lead to unfair identifications; the case of interest here is the kind of parade which consists of people who are highly dissimilar. The measure of facial similarity is consequently only a measure of 'structural fairness', and it should further be noted that facial similarity is only one component of physical similarity. There is a case to be made for the overriding importance of facial similarity, but it may simply be sufficient to note that it is now common police practice in the USA to use photospread lineups, composed only of head-and shoulder

photographs. Facial similarity is just about the only component of physical similarity in this kind of lineup, and the measure developed in this thesis may be justified, if necessary, in application only to photospread lineups.

The legal requirement of physical similarity, at least in South Africa and English law, is that parade members bear a sufficient likeness to the suspect. As I have noted at several places in the thesis, recent publications in the psychological literature reject this requirement, arguing instead that foils should be chosen on the basis of the description given to police at the first stage of the criminal investigation. This contention is not uncontroversial, even in the psychological literature. It must come into conflict with the PC measure of similarity, since the measure assesses 'structural fairness' under the assumption that target-foil similarity is a determinant of this fairness. The task of reconciling the approaches - or, conversely, disposing of one - is not feasible at this stage of the thesis, but it is worth pointing out that the 'match to suspect' strategy must sufficiently meet a 'match to description' strategy, if it is assumed that the suspect matches the description: if the foils resemble the suspect, and the suspect resembles the description, then it follows that the foils must resemble the description.[327]

The final empirical study of the thesis, Study 7, broadened the scope of the empirical enquiry to include questions additional to those concerning the validity and utility of the PC similarity measure. In particular, Study 7 explored the relationship between target-foil similarity and identification accuracy in a simulated identification scenario. The motivation for this study derived from the observation in Chapter 4 that comparatively little is known about this relationship, and that facial similarity may consequently operate as a type of confounding variable. The study was therefore intended as a serious and substantive investigation, but it also served as a demonstration of the potential of the similarity measure as a research tool.

The identification scenario used in Study 7 exposed subjects, in a disguised fashion, to photographs of targets, and later tested their recognition ability with photospread parades. There was no attempt to dress the scenario up as a 'real event' (as, for example, in Malpass & Devine, 1981b, or Murray & Wells, 1982), since my concern was only with recognition accuracy.

There were several important findings in the study. In the first place, identification performance was superior in sequential lineups, but only in target absent lineups, and not in all similarity conditions. The finding of an advantage for sequential parades, though, adds to what is now a solid corpus of findings indicating the superiority of this mode of parade over the traditional mode used by police, at least in South Africa (see Table 4.7). But the study also revealed a negative relation between facial

[327] Of course, it may be 'over-sufficient', in the sense that the match between suspect and foils is too close, making accurate identification too difficult.

similarity and identification performance: lower degrees of similarity led to better identification performance. This was the case in target present lineups, in target absent lineups, and in both simultaneous and sequential lineups.

This latter finding is explicable in some respects, but in other respects it is quite unanticipated. Wells and colleagues (Wells, Rydell & Seelau, 1993; Wells, Seelau, Rydell & Luus, 1994) have recently argued that lineups should have adequate feature heterogeneity - that is suspects and foils should be somewhat dissimilar - since feature homogeneity will impede correct identifications of guilty suspects. In this sense, the finding in question is not surprising: identification decisions were indeed more accurate in low similarity, target-present conditions of Study 7. However, identification decisions were also more accurate in low similarity, target-absent conditions. This is not anticipated: indeed it is directly contrary to an expectation declared by Wells (1984, p 92).

A possible explanation for this finding may lie in the construction of the lineups used in Study 7 (these lineups are shown as Appendix K). The low similarity lineup was constructed by selecting foils whose similarity scores were in the 90 - 100 percentile range, determined in relation to the target (i.e. their spatial distance from the target was great). In the target-present lineups, the target and foils were therefore highly dissimilar. However, in the target-absent lineups, the target was replaced by another foil, drawn again from the pool of foils who fell in the 90 - 100 percentile range. This has the consequence that none of the foils in the target absent lineup resembled the target, and subjects would probably not be drawn easily into making a positive identification decision in such a lineup. If a different selection strategy had been used, say selecting a stand-in foil who was facially similar to the target, the lineup would have had only one foil who resembled the target, and subjects would probably have been enticed with greater frequency into making an identification.

If this explanation holds true, then the finding of an overall recognition advantage for low similarity lineups will only exist for lineups where the suspect bears little resemblance to the perpetrator. At present, there is little information about the resemblance, in real cases, that innocent suspects bear to the perpetrator, and it may be worth treating suspect-perpetrator resemblance as a substantive issue in later research.

The empirical work, in sum, shows that a similarity measure derived from a principal component analysis of face images has considerable potential, and may in addition be a useful research tool. There is much to do, though, in terms of further investigation of the measure, and I wish to draw attention to a few possibilities here.

In the first place - as I pointed out earlier in the Chapter - the technology behind the measure is a bit cumbersome, and would do with some streamlining. The measure needs to be integrated into a

system which has facilities for on-line capture of face images, automation of the required standardization, and relatively effortless derivation of similarity scores.

Although several researchers consider image sets of size 100 adequate for a representational basis of facial images, this is a somewhat dubious claim. We need to investigate the differences between representational bases generated by different populations of faces, before we can have any confidence in the adequacy of a generating set of any particular size. For example, is it better to construct different representational bases for different populations of faces, or to construct one, overarching basis? This is an argument Valentine considers in his theory of the perceptual representation of faces (see Chapter 5), and he shows that there are different implications.

Only one study reported in this thesis investigated the relation of the similarity measure to parade fairness, and it is clear that we need many more before we can claim to understand this relationship. In particular, it will be useful to systematically examine the dependence of indices of lineup fairness on the nature of the written description given to mock witnesses, and the relation of this dependency, in turn, with the measure of facial similarity.

Although Wells and colleagues reject the notion that an 'optimal function' can be determined in respect of target-foil similarity, in part on the basis that it would be difficult to find a suitable 'cut off point' (see the earlier discussion in Chapter 4), I think that the similarity measure may well lend itself to such an investigation. The relation between target-foil similarity and identification accuracy (or perhaps, 'diagnosticity') could be examined empirically, and standard techniques in the differential calculus could be used to find local and global 'maxima'. These will be points at which levels of target-foil similarity produce maximum identification accuracy. There are many complications, of course: we have just seen that suspect-perpetrator similarity is an important additional variable, and there will doubtless be many more.

I started this thesis by asserting that identification research is a type of applied psychology, and that it exists by virtue of a concern with the world of practice. It can only deserve this name, in addition, if it pursues its natural goal, which is application. In this sense, most of what has preceded this chapter in the thesis is only a prelude to real work. I would like, nevertheless, to sketch possible beginnings for such work, and in the last few paragraphs of this thesis I suggest ways in which the facial similarity measure developed here might be put to use. This discussion is necessarily speculative.

Police officials frequently complain about the problems involved in constructing identification parades. Willing foils are difficult to find, and willing foils who bear a reasonable resemblance to the suspect are as scarce as hens' teeth. These problems are less severe when police are allowed to use photospread lineups instead of corporeal parades, but even then the search for a reasonable set of foils can be troublesome. The facial similarity metric could be incorporated into a computerised system for

producing photospreads, and this would make the task of lineup construction much simpler. An image of the suspect could be scanned, and a database of faces could be searched for a number of foils who are within a certain distance from the suspect, which is to be determined from spatial position in the principal component space. Foils selected in this manner could be entered into a 'simultaneous photospread' or 'sequential photospread'.

Particular photospread lineups constructed in the way outlined immediately above - or constructed in other way, actually - could be 'calibrated' on indices of lineup fairness developed in the psychological literature, since Study 6 showed that there is a strong relation between these and the facial similarity measure. However, much more research is required in this respect before the relation is sufficiently well understood for this to be justified.

Just as the similarity metric could underlie a system for constructing photospreads, so it could serve as the heuristic 'engine' of a mugshot retrieval system. Mugshots are presently filed, at best, according to gross physical characteristics such as race, sex and age, and searching through a mugshot album can be a lengthy procedure. Mugshots could be digitally scanned, and then organised - and searched - according to the similarity metric. In particular, a binary search procedure could be implemented: the witness is shown the two most dissimilar images in the database, asked which image is most similar to the perpetrator, and the database is halved according to this choice. The procedure is repeated until the remaining set of images is small enough for the witness to examine entirely. (See Lenorovitz & Laughery, 1984, for a similar suggestion).

Finally, I think that the most powerful potential of the similarity metric and the representational basis underlying it, is to produce synthetic face images. Faces that are members of the space can be reconstructed as images by compositing weighted basis vectors, and this implies that synthetic faces can be constructed in a similar way. An attempt at constructing synthetic (or fictitious) faces was reported in Chapter 8, and although the synthetic faces produced there were not very convincing, they were clearly 'facelike' in appearance. The capability of creating synthetic faces is an interesting one; it may allow Identikit operators and other police artists the opportunity to create much more realistic images.

This potential application of the facial similarity metric may be the most powerful, but it is also the most speculative. Speculation is a significant tool, though, in the prefiguring of psychology in the world.

References

Abdi. H. (1986). Faces, prototypes, and additive tree representations. In H. Ellis, M. Jeeves, F. Newcombe and A. Young. (Eds). *Aspects of face processing*. Pp. 178-84. Dordrecht: Marthinus Nijhoff.

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley and Sons

Agresti, A. & Agresti, B. (1978). Statistical analysis of qualitative variation. In K.F. Schuessler (Ed.). *Sociological Methodology*. Pp. 204-237. San Francisco, CA: Jossey-Bass.

Ainsworth, P.B. & King, E. (1988). Witnesses' perceptions of identification parades. In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds). *Practical Aspects of Memory*. Pp. 67-70. London: John Wiley.

Anastasi, A. (1964). *Fields of Applied Psychology*. New York: McGraw Hill.

Anderson, J.R. & Bower, G.H. (1973). *Human associative memory*. Washington: V.H. Winston.

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology, 3*, 87-104.

Baddeley, A.D. (1979). Applied cognitive and cognitive applied psychology: the case of face recognition. In L.G. Nilsson (Ed.) *Perspectives on memory research*. Hillsdale, N.J.:Erlbaum.

Baddeley, A.D. (1988). But what the hell is it for? In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds). *Practical aspects of memory: vol 1*. Pp. 3-18. London: John Wiley.

Bahrick, H.P., Bahrick, P.G. & Wittlinger, R.P. (1975). Fifty years of memory for names and faces. *Journal of Experimental Psychology: General, 104*, 54-75.

Bahrick, H.P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113*, 1-29.

Banaji, M.R. & Crowder, R.G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44*(9), 1185-93.

Bartlett, J.C. & Hurry, S. (1984). Typicality and familiarity of faces. *Memory & Cognition, 12*(3), 219-228.

Bechtel, W. (1991). *Connectionism and the mind*. Cambridge, Mass.: Basil Blackwell.

Bekerian, D. A. (1993). In search of the typical eyewitness. *American Psychologist, 48*, 574-6.

Benson, P.J. & Perrett, D.I. (1991a). Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology, 3*(1), 105-135.

Benson, P.J. & Perrett, D.I. (1991b). Synthesizing continuous tone caricatures. *Images & Vision Computing, 9*(2), 123-129.

Benson, P.J., Perrett, D.I. & Davis, D.N. (1992). Towards a quantitative understanding of facial caricatures. In V. Bruce & A.M. Burton (Eds). *Processing Images of Faces*. N.J.: Ablex.

Berkowitz, L. & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37*(3), 245-257.

Bloom, L.C. & Mudd, S.A. (1991). Depth of processing approach to face recognition: A test of two theories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*(3), 556-65.

Bowns, L. & Morgan, M.J. (1993). Facial features and axis of symmetry extracted using natural orientation information. *Biological Cybernetics, 70*(2), 137-44.

Bradshaw, J.L. & Wallace, G. (1971). Models for the processing and identification of faces. *Perception & Psychophysics, 9*(5), 443-8.

Brennan, S.E. (1985). The caricature generator. *Leonardo, 18*, 170-178.

Brigham, J.C. & Brandt, C.C. (1992). Measuring lineup fairness: mock witness responses versus direct evaluations of lineups. *Law and Human Behaviour, 16*(5), 475-489.

Brigham, J.C. & Cairns, D.L. (1988). The effect of mugshot inspections on eyewitness identification accuracy. *Journal of Applied Social Psychology, 18*(16), 1394-1410.

Brigham, J.C. & Malpass, R.S. (1985). The role of experience and contact in the recognition of faces of own- and other-race persons. *Journal of Social Issues, 41*(3), 139-155.

Brigham, J.C. & Pfeifer, J.E. (1994). Evaluating the fairness of lineups. In D. Ross, J.D. Read, & M.P. Toglia (Eds). *Adult eyewitness testimony: Current trends and developments*. Pp. 201-222. New York: Cambridge University Press.

Brigham, J.C. & Ready, D.J. (1985). Own-race bias in lineup construction. *Law and Human Behavior, 9*, 415-24.

Brooks, N. (1983). *Pre-trial eyewitness identification procedures: Police guidelines*. Ottawa: Law Reform Commission of Canada.

Bruce, V. (1979). Searching for politicians: An information-processing approach to face recognition. *Quarterly Journal of Experimental Psychology, 31*, 373-95.

Bruce, V. (1988). *Recognizing faces*. Sussex, U.K.: Lawrence Erlbaum.

Bruce, V. (1992). Perceiving and recognizing faces. In G.W. Humphreys (Ed). *Understanding vision*. Pp. 87-103. Oxford: Basil Blackwell.

Bruce, V. (1993). Commentary: Dimensions of facial appearance. In G.M. Davies & R.H. Logie (Eds).

Memory in everyday life. Pp 351-9. Elsevier Science Publishers.

Bruce, V. (1994). Stability from variation: The case of face recognition. *Quarterly Journal of Experimental Psychology: Human, 47A*(1), 5-28.

Bruce, V. & Burton, M. (1989). Computer recognition of faces. In A.W. Young and H.D. Ellis (Eds). *Handbook of research on face processing*. 487-506. North Holland: Elsevier Science Publishers.

Bruce, V. & Valentine, T. (1988). When a nod's as good as a wink: The role of dynamic information in facial recognition. In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds). *Practical aspects of memory*. Pp. 169-174. London: John Wiley.

Bruce, V. & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305-327.

Bruce, V., Burton, A.M. & Craw, I. (1992). Modelling face recognition. *Philosophical Transactions of the Royal Society of London, B, 335*, 121-128.

Bruce, V., Burton, A.M. & Dench, N. (1994). What's distinctive about a distinctive face? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 47A*(1), 119-41.

Bruce, V., Burton, A.M. & Walker, S. (1994). Testing the models? New data and commentary on Stanhope and Cohen (1993). *British Journal of Psychology*, 85, 335-49.

Bruce, V., Burton, M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R. & Linney, A. (1993). Sex discrimination: how do we tell the difference between male and female faces? *Perception, 22*, 131-52.

Bruce, V., Coombes, A. & Richards, R. (1993). Describing the shapes of faces using surface primitives. *Image and Vision Computing, 11*(6), 353-63.

Bruce, V., Doyle, T., Dench, N. & Burton, M. (1991). Remembering facial configurations. *Cognition*, 38(2), 109-144.

Bruce, V., Hanna, E., Dench, N., Healey, P. & Burton, M. (1992). The importance of mass in line drawings of faces. *Applied Cognitive Psychology, 6*, 619-628.

Bruce, V., Healey, P. (1991). Recognising facial surfaces. *Perception, 20*(6) 755-769.

Brunelli, R. & Poggio, T. (1993). Caricatural effects in automated face perception. *Biological Cybernetics, 69*(3), 235-41.

Buckhout, R. (1974). Eyewitness testimony. *Scientific American, 231*(12), 23-31.

Buckhout, R., Figuero, D. & Hoff, E. (1975). Eyewitness identification: Effects of suggestion and bias in identification from photographs. *Bulletin of the Psychonomic Society, 6*(1), 71-74.

Buckhout, R., Rabinowitz, M., Alfonso, V., Kanellis, D. & Anderson, J. (1988). Empirical assessment of lineups: Getting down to cases. *Law and Human Behaviour, 12*(3), 323-331.

Bull, R. & Clifford, B.R. (1984). Earwitness voice recognition accuracy. In G.L. Wells & E.F. Loftus (Eds). *Eyewitness testimony: Psychological perspectives*. Pp 92-123. Cambridge: Cambridge University Press.

Burton, A.M. (1992). Good morning, Mr . . . er: why do we recognise a face, but sometimes draw a blank when it comes to the name? *New Scientist, 133* (Feb. 1 '92), 39-41.

Burton, A.M. & Bruce, V. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*(3), 361-380.

Burton, A.M., Bruce, V. & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception, 22*, 153-76.

Calis, G.J., Sterenborg, J. & Maarse, F. (1984). Initial microgenetic steps in single-glance face recognition. *Acta Psychologica, 55*(3), 215-230.

Campos, J.C., Linney, A.D. & Moss, J.P. (1993). The analysis of facial profiles using scale-space techniques. *Pattern Recognition, 26*(6), 819-24.

Carey, S. (1992). Becoming a face expert. *Philosophical Transactions of the Royal Society of London, B, 335*, 95-103.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-76.

Cheng, Y.Q., Liu, K., & Yang, J.Y. (1993). A novel feature-extraction method for image recognition based on similar discriminant function (Sdf). *Pattern Recognition, 26*(6), 115-25.

Cohen, G. (1990). *Memory in the real world*. Sussex: Lawrence Erlbaum.

Coombes, A.M., Moss, J.P., Linney, A.D., Richards, R. & Jasmes, D.R. (1991). A mathematical method for the comparison of three-dimensional changes in the facial surface. *European Journal of Orthodontics, 13*, 95-110.

Craw, I. & Cameron, P. (1991). Parameterising images for recognition and reconstruction. *Proceedings of the British Machine Vision Conference BMCV '91*. Turing Institute Press and Springer Verlag.

Cunningham, M.R., Barbee, A.P. & Pike, C.L. (1990). What do women want? Facialmetric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of Personality and Social Psychology, 59*(1), 61-72.

Cutler, B. & Penrod, S. (1988). Improving the reliability of eyewitness identification: lineup construction and presentation. *Journal of Applied Psychology, 73*(2), 281-90.

Cutler, B.L., Berman, G.L., Penrod, S. & Fisher, R.P. (1994). Conceptual, practical and empirical issues associated with eyewitness identification test media. In D. Ross, J.D. Read, & M.P. Toglia (Eds). *Adult Eyewitness testimony: Current trends and developments*. Pp. 163-81. New York: Cambridge University Press.

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 Behaviour, 11*(3), 233-58.

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*(4), 629-37.

Danziger, K. (1990). The autonomy of applied psychology. Paper presented at *22nd International Congress of Applied Psychology*, Kyoto, July, 1990.

Davidoff, J. (1986). The mental representation of faces: Spatial and temporal factors. *Perception & Psychophysics, 40*(6), 391-400.

Davies, G.M. (1988). Faces and places: Laboratory research on context and face recognition. In G.M. Davies & D. Thomson (Eds). *Memory in context: Context in memory*. Pp 35-53. Chichester: John Wiley & Sons.

Davies, G.M. (1989). The applicability of facial memory research. In A.W. Young & H.D. Ellis (Eds). *Handbook of research on face processing*. Pp. 557-61. North Holland: Elsevier Science Publishers.

Davies, G.M. (1992). Influencing public policy on eyewitnessing: Problems and possibilities. In F. Losel, T. Blisener & D. Beider (Eds). *Psychology and law: International perspectives*. Pp. 265-74. Berlin: De Greuter.

Davies, G. M. (1993). Witnessing events. In G.M. Davies & R.H. Logie (Eds). *Memory in everyday life*. Pp 367-401. Elsevier Science Publishers.

Davies, G. & Milne, A. (1982). Recognizing faces in and out of context. *Current Psychological Research, 2*(4), 235-246.

Davies, G.M., Shepherd. J.W. & Ellis, H.D. (1979). Similarity effects in face recognition. *American Journal of Psychology, 92*, 507-23.

Deffenbacher, K. (1991). A maturing of research on the behaviour of eyewitnesses. *Applied Cognitive Psychology*, 5, 377-402.

Dent, H.R. (1977). Stress as a factor influencing person recognition in identification parades. *Bulletin of the British Psychological Society, 30*, 339-40.

Devine, P.G. & Malpass, R.S. (1985). Orienting strategies in differential face recognition. *Personality and Social Psychology Bulletin, 11*(1), 33-40.

Devlin, Lord Patrick. (1976). *Report to the Secretary of State for the Home Department of the Departmental Committee on Evidence of Identification in Criminal Cases*. London: Her Majesty's Stationery Office.

Dewdney, A.K. (1986). Computer recreations. *Scientific American, 255* (October), 20-27.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107-17.

DiNardo, L. & Rainey, D.W. (1989). Recognizing faces in bright and dim light. *Perceptual & Motor Skills, 68*(3), 836-838.

Doob, A.N. (1980). Police procedures and psychological knowledge in eyewitness identification. Paper presented at *the Alberta Eyewitness Conference, Edmonton, Alberta.*

Doob, A.N. & Kirshenbaum, H.M. (1973). Bias in police lineups - partial remembering. *Journal of Police Science and Administration, 1*(3), 287-93.

Dudycha, G. (1963). *Applied Psychology*. New York: Rank Press Co.

Ebbinghaus. H.E. (1964). *Memory: A contribution to experimental psychology*. New York: Dover.

Egan, D., Pittner, M. & Goldstein, A.G. (1977). Eyewitness identification: Photographs vs live models. *Law and Human Behaviour, 1*(2), 199-206.

Egeth, H.E. (1993). What do we not know about eyewitness identification? *American Psychologist, 48*, 577-80.

Ellis, A.W. (1992). Cognitive mechanisms of face processing. *Philosophical Transactions of the Royal Society of London, B, 335*, 113-119.

Ellis, A.W., Young, A.W., Flude, B.M. & Hay, D.C. (1991). Repetition priming of face recognition. *Quarterly Journal of Experimental Psychology: Human, 39A*, 193-210.

Ellis, H.D. & Shepherd, J.W. (1992). Face memory - theory and practice. In M.M. Gruneberg & P.E. Morris, (Eds). *Aspects of memory*. Pp 1-17. London: Routledge.

Ellis, H.D. (1984). Practical aspects of face memory. In G.L. Wells & E.F. Loftus (Eds). *Eyewitness testimony: Psychological perspectives*. Pp 12-37. Cambridge: Cambridge University Press.

Ellis, H.D. (1992). The development of face processing skills. *Philosophical Transactions of the Royal Society of London, B, 335*, 105-111.

Ellis, H.D., Shepherd, J.W. & Davies, G.M. (1979). Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. *Perception, 8*, 431-9.

Ellis, H.D., Shepherd, J.W., Shepherd, J., Flin, R. & Davies, G.M. (1989). Identification from a computer-driven retrieval system compared with a traditional mug-shot album search: A new tool for police investigations. *Ergonomics, 32*(2), 167-177.

Ellison, K.W. & Buckhout, R. (1981). *Psychology and criminal justice*. New York: Harper & Row.

Everitt, B.S. & Dunn, G. (1991). *Applied multivariate data analysis*. London: Edward Arnold.

Flury, B. & Riedwyl, H. (1988). *Mutlivariate Statistics: A practical approach*. London: Chapman & Hall.

Gibson, E. (1969). *Principles of perceptual learning and development*. New York: Appleton Century Crofts.

Goldstein, A.G., Chance, J.E. & Schneller, G.R. (1989). Frequency of eyewitness identification in criminal cases: a survey of prosecutors. *Bulletin of the Psychonomic Society, 27*(1), 71-74.

Gonzalez, R., Ellsworth, P. & Pembroke, M. (1993). Response bias in lineups and showups. *Journal of Personality and Social Psychology, 64*(4), 525-37.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40,* 237-64.

Goodman, J. & Loftus, E. (1989). Implications of facial memory research for investigative and administrative criminal procedures. In A.W. Young & H.D. Ellis (Eds). *Handbook of research on face processing.* Pp. 571-8. North Holland: Elsevier Science Publishers.

Gorenstein, G.W. & Ellsworth, P.C. (1980). Effect of choosing an incorrect photograph on a later identification by an eyewitness. *Journal of Applied Psychology, 65*, 616-622.

Green, D.L. & Geiselman, R. E. (1989). Building composite facial images: Effects of feature saliency and delay of construction *Journal of Applied Psychology, 74*, 714-21.

Gruneberg, M.M. & Morris, P.E. (1992). Applying memory research. In M.M. Gruneberg & P.E. Morris, (Eds). *Aspects of memory.* Pp 1-17. London: Routledge.

Hagen, M.A. & Perkins, D. (1983). A refutation of the hypothesis of the superfidelity of caricatures relative to photographs. *Perception, 12*(1) 55-61.

Haig, N.D. (1984). The effect of feature displacement on face recognition. *Perception, 13*(5) 505-512.

Haig, N.D. (1986a). Exploring recognition with interchanged facial features. *Perception, 15*(3) 235-247.

Haig, N.D. (1986b). High-resolution facial feature saliency mapping. *Perception, 15*(4), 373-386.

Hancock, P.J.B., Burton, A.M. & Bruce, V. (1994). Face processing: human perception and principal components analysis. Submitted for publication to *Memory and Cognition.*

Harmon, L.D. (1973). The recognition of faces. *Scientific American, 229*, 70-82.

Harries, M.H. & Perrett, D.I. (1991). Preferential inspection of views of 3-D model heads. *Perception, 20*(5), 669-680.

Hay, D.C. & Young, A.W. (1982). The human face. In A.W. Ellis (eds.), *Normality and pathology in cognitive functions.* London: Academic Press.

Hay, D.C. & Young, A.W. (1991). Routes through the face recognition system. *Quarterly Journal of Experimental Psychology: Human, 43A*(4), 761-91.

Hays, W.L. (1994). *Statistics.* 5th edition. New York: Harcourt Brace.

Hirschberg, N., Jones, L.E. & Haggerty, M. (1978). What's in a face: individual differences in face perception. *Journal of Research in Personality, 12*, 488-99.

Hoffman, L. & Zefertt, D.T. (1989). *South African Law of Evidence.* Durban: Butterworths

Honeck, R.P. (1986). A serendipitous finding in face recognition. *Bulletin of the Psychonomic Society, 24*(5), 369-371.

Howell, D.C. (1992). *Statistical Methods for Psychology.* 3rd edition. Boston: PWS Kent.

Huang, C.L. & Chen, C.W. (1992). Human facial feature-extraction for face interpretation and recognition. *Pattern Recognition, 25*(12), 115-25.

Ihde, D. (1979). *Technics and Praxis: Boston studies in the philosophy of science*, vol xxiv. Boston: D. Reidel Publishing company.

Inui, T. & Miyamoto, K. (1984). The effect of changes in visible area on facial recognition. *Perception, 13*(1) 49-56.

Jenkins, F. & Davies, G.M. (1985). Contamination of facial memory through exposure to misleading composite pictures (photofit). *Journal of Applied Psychology, 70,* 164-76.

Jensen, D.G. (1987). Facial perception studies using the Macintosh. *Behavior Research Methods, Instruments and Computers, 19*(2), 252-6.

Johnston, B. (1993). Commentary: Accessing identity information. In G.M. Davies & R.H. Logie (Eds). *Memory in everyday life.* Pp 360-6. Elsevier Science Publishers.

Joliffe, I.T. (1972). Discarding variables in a principal components analysis I: Artificial data. *Applied Statistics, 21*, 160-73.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187-200.

Kassin, S.M., Ellsworth, P.C. & Smith, V.L. (1989). The "general acceptance" of psychological research on eyewitness testimony. *American Psychologist, 44*(8), 1089-98.

Kennedy, J.J. (1992). *Analyzing qualitative data.* 2nd ed. New York: Praeger.

Kirby, M. & Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterisation of human faces. *IEEE: Transactions on Pattern Analysis and Machine Intelligence, 12*, 103-8.

Klatzky, R.L. (1980). *Human memory: structures and processes.* San Francisco: W.H. Freeman.

Klatzky, R.L. (1991). Let's be friends. *American Psychologist, 46*(1), 43-45.

Klatzky, R.L. & Forrest, F.H. (1984). Recognizing familiar and unfamiliar faces. *Memory & Cognition, 12*(1), 60-70.

Köhnken, G. & Maass, A. (1988). Eyewitness testimony: False alarms on biased instructions? *Journal of Applied Psychology, 73*(3), 363-70.

Kohonen, T. (1984). *Self-organization and associative memory.* Berlin: Springer-Verlag.

Konecni, V.J. & Ebbesen, E.B. (1986). Courtroom testimony by psychologists on eyewitness identification issues. *Law and Human Behaviour, 10*(1/2), 117-26.

Kreutzer, M.A., Leonard, C. & Flavell, J. (1975). An interview study of children's knowledge about

memory. *Monographs of the Society for research in child development, 40*, Serial No. 159.

Krumhansl, C.L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*(5), 445-63.

Lang, S. (1987). *Linear algebra.* Connecticut, USA: Springer Verlag.

Langlois, J.H. & Roggman, L.A. (1990). Attractive faces are only average. *Psychological Science, 1*(2) 115-121.

Langlois, J.H. & Roggman, L.A. (1991). A picture is worth a thousand words: Reply to "On the difficulty of averaging faces." *Psychological Science, 2*(5) 354-357.

Laughery, K.R. & Wogalter, M.S. (1989). Forensic applications of facial memory research. In A.W. Young & H.D. Ellis (Eds). *Handbook of research on face processing.* Pp. 519-48. North Holland: Elsevier Science Publishers.

Laughery, K.R., Fessler, P.K., Lenorovitz, D.R. & Yoblick, D.A. (1974). Time delay and similarity effects in facial recognition. *Journal of Applied Psychology, 39*(4), 490-6.

Laughery, K.R., Jensen, D.G. & Wogalter, M.S. (1988). Response bias with prototypic faces. In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds). *Practical aspects of memory.* Pp. 157-162. London: John Wiley.

Leippe, M.R. (1985). The influence of eyewitness non-identifications on mock-jurors' judgements of a court case. *Journal of Applied Social Psychology, 15*(7), 656-660.

Lenorovitz, D.R. & Laughery, K.R. (1984). A witness-computer interactive system for searching mug files. In G.L. Wells & E.F. Loftus (Eds). *Eyewitness testimony: Psychological perspectives.* Pp 38-63. Cambridge: Cambridge University Press.

Li, H.Y., Qiao, Y. & Psaltis, D. (1993). Optical network for real-time face recognition. *Applied Optics, 32*(26), 5026-35.

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. (1986). Confidence and accuracy of eyewitness identification from lineups. *Law and Human Behaviour, 10*(3), 229-39.

Lindsay, R.C.L. (1994). Biased lineups: Where do they come from? In D. Ross, J.D. Read, & M.P. Toglia (Eds). *Adult eyewitness testimony: Current trends and developments.* Pp. 182-200. New York: Cambridge University Press.

Lindsay, R.C.L. & Wells, G.L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behaviour, 4*(4), 303-313

Lindsay, R.C.L. & Wells, G.L. (1985). Improving eyewitness identifications from lineups: simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556-564.

Lindsay, R.C.L., Lea, J.A., Nosworthy, G.J., Fulford, J.A., Hector, J., LeVan, V & Seabrook, C. (1991). Biased lineups: sequential presentation reduces the problem. *Journal of Applied Psychology, 76*(6), 796-802.

Lindsay, R.C.L., Lea, J.G. & Fulford, J.A. (1991). Sequential lineup presentation: technique matters. *Journal of Applied Psychology, 76*(5), 741-45.

Lindsay, R.C.L., Nosworthy, G.J., Martin, R. & Martynuck, C. (1994). Using mug shots to find suspects. *Journal of Applied Psychology, 79*(1), 121-30.

Lindsay, R.C.L., Wallbridge, H. & Drennan, D. (1987). Do the clothes make the man?: An exploration of the effect of lineup attire on eyewitness identification accuracy. *Canadian Journal of Behavioral Science, 19*(4), 463-78.

Liu, K., Cheng, Y.Q. & Yang, J.Y. (1993). Algebraic feature-extraction for image recognition based on an optimal discriminant criterion. *Pattern Recognition, 26*(6), 903-11.

Lloyd-Bostock, S. & Clifford, B. (1983). *Evaluating witness evidence.* London: John Wiley and Sons.

Loftus, E.F. (1979). *Eyewitness testimony.* Harvard, Mass.: Harvard University Press.

Loftus, E.F. (1983a). Silence is not golden. *American Psychologist, 38*(5), 564-72.

Loftus, E.F. (1983b). Whose shadow is crooked? *American Psychologist, 38*(5), 576-77.

Loftus, E.F. (1993). Commentary: The theory behind witnessing events, and the practice. In G.M. Davies & R.H. Logie (Eds). *Memory in everyday life.* Pp 402-7. Elsevier Science Publishers.

Loftus, E.F. & Greene, E. (1980). Warning: Even memory for faces may be contagious. *Law and Human Behavior, 4*, 323-334.

Loftus, E.F. & Ketcham, K. (1991). *Witness for the defence.* New York: St. Martin's Press.

Logie, R.H. & Baddeley, A.D. (1987). Face recognition, pose and ecological validity. *Applied Cognitive Psychology, 1*(1), 53-69.

Luukkonen, T. & Stahle, B. (1990). Qualitative evaluations in management of basic and applied research. *Research Policy, 19*(4), 357-68.

Luus, E.C.A. & Wells, G.L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behaviour, 15*(1), 43-57.

Malpass, R.S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behaviour, 5*(4), 299-309.

Malpass, R.S. & Devine, P.G. (1981a). Eyewitness identification: lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66*(4), 482-489.

Malpass, R.S. & Devine, P.G. (1981b). Realism and eyewitness identification research. *Law and Human Behavior, 4*(4), 347-58.

Malpass, R.S. & Devine, P.G. (1983). Measuring the fairness of eyewitness identification lineups. In S.M.A. Lloyd-Bostock & B.R. Clifford. *Evaluating Witness Evidence*. London: John Wiley and sons.

Malpass, R.S. & Devine, P.G. (1984). Research on suggestion in lineups and photospreads. G.L. Wells & E.F. Loftus (Eds). *Eyewitness testimony: Psychological perspectives*. Pp 12-37. Cambridge: Cambridge University Press.

Malpass, R.S. & Hughes, K.D. (1986). Formation of facial prototypes. In H. Ellis, M. Jeeves, F. Newcombe & A. Young. (Eds). *Aspects of face processing*. Pp. 154-62. Dordrecht: Marthinus Nijhoff.

Malpass, R.S. & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*, 330-334.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Mauro, R. & Kubovy, M. (1992). Caricature and face recognition. Special issue: Memory and cognition applied. *Memory & Cognition*, *20*(4), 433-440.

McAllister, H.A. & Bregman, N.J. (1986). Juror underutilization of eyewitness nonidentifications: theoretical and practical implications *Journal of Applied Psychology, 71*, 168-70.

McAllister, H.A. & Bregman, N.J. (1989). Juror underutilization of eyewitness nonidentifications: a test of the disconfirmed expectancy explanation. *Journal of Applied Social Psychology*, *19*(1), 20-29.

McAllister, H.A., Dale, R.H.I. & Keay, C.E. (1993). Effects of lineup modality on witness credibility *The Journal of Social Psychology*, *133*, 365-76.

McClelland, J.L. & Rumelhart, D.E. (1981). An interactive activation model of the effect of context in perception. *Psychological Review*, *88*, 375-406.

McClelland, J.L. & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159-188.

McClelland, J.L. & Rumelhart, D.E. (1985). *Parallel distributed processing* (Vol. 2). Cambridge, MA: MIT Press.

McCloskey, M. & Egeth, H. (1983). Eyewitness identification: what can a psychologist tell a jury? *American Psychologist*, *38*, 550-563.

McKelvie, S.J. (1987). Recognition memory for faces with and without spectacles. *Perceptual & Motor Skills*, *65*(3), 705-706.

McKelvie, S.J. (1991). Effects of processing strategy and transformation on recognition memory for photographs of faces. *Bulletin of the Psychonomic Society*, *29*(2), 98-100.

Melara, R.D., De Witt-Rickards, T. & O'Brien, T.P. (1989). Enhancing lineup identification accuracy: two codes are better than one. *Journal of Applied Psychology*, *74*(5), 706-13.

Melton, G. (1987). Bringing psychology to the legal system. *American Psychologist*, *42*, 488-95.

Memon, A., Dionne, R., Short, L., Maralani, S., MacKinnon, D. & Geiselman, R.E. (1988). Psychological factors in the use of photospreads. *Journal of Police Science and Administration*, *16*(1), 62-69.

Milord, J.T. (1978). Aesthetic aspects of faces: A (somewhat) phenomenological analysis using multidimensional scaling methods. *Journal of Personality and Social Psychology, 36*(2), 205-16.

Monahan, J. and Walker, L. (1988). Social science research in law: A new paradigm. *American Psychologist, 43*, 465-72.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165-178.

Morton, J. (1991). The bankruptcy of everyday thinking. *American Psychologist*, *46*(1), 32-33.

Mumford, D. (1991). Parameterizing exemplars of categories. *Journal of Cognitive Neuroscience*, *3*(1), 87-8.

Münsterberg, H. (1908). *On the witness stand: Essays on psychology and crime*. New York: Clark Boardman.

Murray, C., Sloth-Nielsen, J, and Tredoux, C.G. (1989). The death penalty in the Cape Provincial Division 1986 - 1988. *South African Journal on Human Rights*, *5*(2), 154-182.

Murray, D.M. & Wells, G.L. (1982). Does knowledge that a crime was staged affect eyewitness performance? *Journal of Applied Social Psychology*, *12*(1), 42-53.

Navon, D. (1990a). How critical is the accuracy of an eyewitness's memory? Another look at the issue of lineup diagnosticity. *Journal of Applied Psychology, 75*(5), 506-10.

Navon, D. (1990b). Ecological parameters ≠ nonlineup evidence: a reply to Wells and Luus. *Journal of Applied Psychology, 75*(5), 506-10.

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behavior*, *16*, 575-93.

Neisser, U. (1978). Memory: What are the important questions? In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds). *Practical aspects of memory*. London: Academic Press.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton Century Crofts.

Neisser, U. (1976). *Cognition and reality*. San Francisco: Freeman.

Neisser, U. (1983). *Memory observed*. San Francisco: Freeman.

Neisser, U. (1991). A case of misplaced nostalgia. *American Psychologist*, *46*(1), 34-36.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, *17*(1), 3-27.

Nosworthy, G.J. & Lindsay, R.C.L. (1990). Does nominal lineup size matter? *Journal of Applied Psychology, 75*(3), 358-61.

O'Toole, A.J., Milward, R.B. & Anderson, J.A. (1988). A physical system approach to recognition memory for spatially transformed faces. *Neural Networks, 1*, 179-99.

O'Toole, A.J. & Thompson, J.L. (1993). An X Windows tool for synthesizing face images from eigenvectors. *Behavior Research Methods, Instruments and Computers, 25*(1), 41-7.

O'Toole, A.J., Abdi, H., Deffenbacher, K.A. & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America, 10*(3), 405-11.

O'Toole, A.J., Deffenbacher, K.A., Valentin. D. & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory and Cognition, 22*(2), 208-24.

Orne, M.T. (1962). On the social psychology of the psychological experiment. *American Psychologist, 17*, 776-783.

Paley, B. & Geiselman, R.E. (1989). The effects of alternative photospread instructions on suspect identification performance. *American Journal of Forensic Psychology, 7*(1), 3-13.

Parker, I. (1989). *The crisis in modern social psychology and how to end it.* London: Routledge.

Patterson, K.E. & Baddeley, A.D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 406-17.

Pearson, D. (1992). The extraction and use of facial features in low bit-rate visual communication. *Philosophical Transactions of the Royal Society of London, B, 335*, 79-85.

Pearson, D., Hanna, E. & Martinez, K. (1990). Computer-generated cartoons. In H. Barlow, C. Blakemore & M. Weston-Smith (Eds). *Images and understanding.* Cambridge University Press.

Pearson, D.E. & Robinson, J.A. (1985). Visual communication at very low data-rates. *Proceedings of the IEEE: Transactions on Pattern Analysis and Machine Intelligence, 73*, 795-811.

Perrett, D.I., Hietanen, J.K., Oram, M.W. & Benson, P.J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London, B, 335*, 23-30.

Phillips, R.J. (1972). Why are faces hard to recognise in photographic negative? *Perception and Psychophysics, 12*, 425-26.

Phillips, W.A. & Smith, L.S. (1989). Conventional and connectionist approaches to face processing by computer. In A.W. Young & H.D. Ellis (Eds). *Handbook of research on face processing.* Pp. 513-20. North Holland: Elsevier Science Publishers.

Pigott, M. & Brigham, J.C. (1985). Relationship between accuracy of prior description and facial recognition *Journal of Applied Psychology, 70*, 547-55.

Pittenger, J.B. (1991). On the difficulty of averaging faces: Comments on Langlois and Roggman. *Psychological Science, 2*(5), 351-353.

Potter, J. (1982). "...nothing so practical as a good theory". The problematic application of social psychology. In P. Stringer (Ed.) *European perspectives on social psychology.* London: Academic Press.

Rakover, S.S. & Cahlon, B. (1989). To catch a thief with a recognition test: The model and some empirical results. *Cognitive Psychology, 21*(4), 423-468.

Read, J.D. & Bruce, D. (1984). On the external validity of questioning effects in eyewitness testimony. *International Review of Applied Psychology, 33*, 33-50.

Read, J.D., Hammersley, R.H., Cross-Calvert, S. & McFadzen, E. (1989). Rehearsal of faces and details in action events. *Applied Cognitive Psychology, 3*, 295-311.

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*(5), 870-82.

Rhodes, G. (1988). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception, 17*, 43-63.

Rhodes, G. (1993a). Configural coding, expertise, and the right-hemisphere advantage for face recognition. *Brain and Cognition, 22*(1), 19-41.

Rhodes, G. (1993b). When do caricatures look good? *New Zealand Journal of Psychology, 22*, 110-113.

Rhodes, G. & MacLean, I.G. (1990). Distinctiveness and expertise effects with homogeneous stimuli: Towards a model of configural coding. *Perception, 19*, 773-94.

Rhodes, G. & Moody, J. (1990). Memory representations of unfamiliar faces: Coding of distinctive information. *New Zealand Journal of Psychology, 19*(2), 70-78.

Rhodes, G. & Wooding, R. (1989). Laterality effects in identification of caricatures and photographs of famous faces. *Brain & Cognition, 9*(2), 201-209.

Rhodes, G., Brake, S. & Taylor, K. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology, 80*, 313-31.

Rhodes, G., Brennan, S. & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19*, 473-97.

Roberts, T. & Bruce, V. (1988). Feature saliency in judging the sex and familiarity of faces. *Perception, 17*(4), 475-481.

Rosch, E. & Mervis, C.B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 1-19.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton Century Crofts.

Rothman, K.J. (1986). *Modern epidemiology*. Boston: Little, Brown and Company.

Rumelhart, D.E. & McClelland, J.L. (1985). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Schiff, W., Banka, L. & De Bordes Galdi, G. (1986). Recognizing people seen in events via dynamic "mug shots." *American Journal of Psychology*, *99*, 219-31.

Schreiber, A. & Rousset, S. (1991). Facenet: A connectionist model of face identification in context. *European Journal of Cognitive Psychology*, *3*(1), 177-98.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology*, *75*, 221-42.

Séve, L. (1976). *Marxism and the theory of human personality*. London: Lawrence and Wishart.

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

Shephard, R.N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*(6), 509-23.

Shephard, R.N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shepherd, J.W. & Ellis, H.D. (1973). The effect of attractiveness on recognition memory for faces. *American Journal of Psychology*, *86*, 627-33.

Shepherd, J.W., Davies, G.M. & Ellis, H. (1981). Studies of cue saliency. In G.M. Davies, H. Ellis, & J. Shepherd. (Eds). *Perceiving and remembering faces*. Pp. 105-31. London: Academic Press.

Shepherd, J.W., Ellis, H. & Davies, G.M. (1982). *Identification evidence: a psychological evaluation*. Aberdeen: Aberdeen University Press.

Shepherd, J.W., Gibling, F. & Ellis, H.D (1991). The effects of distinctiveness, presentation time and delay on face recognition. *European Journal of Cognitive Psychology*, *3*(1), 137-45.

Sirovich, L. & Kirby, M. (1987). Low dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, *4*, 519-24.

Sobel, N.R. & Prigden, D. (1982). *Eyewitness identification: legal and practical problems*. New York: Clark Boardman.

Solso, R. & McCarthy, J. (1981). Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, *72*, 499-503.

Sporer, S.L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*(1), 22-33.

Stehr, N. (1992). *Practical knowledge: Applying the social sciences*. London: Sage.

Stern, L.W. (1910). Abstracts of lectures on the psychology of testimony on the study of individuality. *American Journal of Psychology*, *21*, 270-282.

Stolz, J.B. (1981). Adoptions of innovations from applied behavioural research. *Journal of Applied Behavioural Analysis*, *14*, 491-505.

Stroh, R.C. (1991). Developing sponsors for applied research. *SRA Journal*, *23*(3), 33-40.

Tajfel, H. (1972). Experiments in a vacuum. In J. Israel & H. Tajfel. (eds). *The contexts of Social Psychology: a critical assessment*. London: Academic Press.

Tremper, C. (1987). Organized psychology's efforts to influence judicial decision making. *American Psychologist*, *42*, 496-501.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71-86.

Tversky, B. & Baratz, D. (1985). Memory for faces: Are caricatures better than photographs. *Memory and Cognition*, *13*(1), 45-49.

Valentine, T. (1991a). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, *43A*(2), 161-204.

Valentine, T. (1991b). Representation and process in face recognition. In R. Watt, (Ed.) *Vision and visual dysfunction. Vol 14: Pattern recognition in man and machine*. London: MacMillan.

Valentine, T, Bredart, S., Lawson, R. & Ward, G. (1991). What's in a name? Access to information from people's names. *European Journal of Cognitive Psychology*, *3*(1), 147-76.

Valentine, T. & Bruce, V. (1986a). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology 40*(3), 300-305.

Valentine, T. & Bruce, V. (1986b). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica*, *61*(3), 259-273.

Valentine, T. & Bruce, V. (1986c). The effects of distinctiveness in recognising and classifying faces. *Perception 15*(5) 525-535.

Valentine, T. & Bruce, V. (1988). Mental rotation of faces. *Memory & Cognition*, *16*(6), 556-566.

Valentine, T. & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology: Human*, *44A*(4), 671-703.

Valentine, T. & Ferrara, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology*, *82*(1), 87-102.

van der Vlist, J. (1982). Social psychological theories and empirical study of behavioural problems. In Stringer, P. (Ed). (1982). Confronting Social Issues. *European Monographs in Social Psychology*.

Vannier, M.W., Pilgram, T., Bhatia, G. & Brunsden, B. (1991). Facial surface scanner. *IEEE Computer Graphics and Applications*, November 1991, 72-80.

Vokey, J.R. & Read, J.D. (1988). Typicality, familiarity and the recognition of male and female faces. *Canadian Journal of Psychology, 42*, 489-95.

Vokey, J.R. & Read, J.D. (1992). Familiarity, memorability and the effect of typicality on the recognition of faces. *Memory and Cognition, 20*(3), *291-302.*

Wagenaar, W.A. & Veefkind, N. (1992). Comparison of one-person and many-person lineups: A warning against unsafe practices. In F. Losel, T. Blisener & D. Beider. (Eds). *Psychology and law: International perspectives*. Pp. 275-85. Berlin: De Greuter.

Warr, P. (1978). *Psychology at Work* Harmondsworth: Penguin.

Wells, G.L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*, 1546-57.

Wells, G.L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, 14(2), 89-103.

Wells, G.L. (1985). Verbal descriptions of faces from memory: Are they diagnostic of identification accuracy? *Journal of Applied Psychology, 70*(4), 619-626.

Wells, G.L. (1986). Practical issues in eyewitness research. In M.F. Kaplan (Ed.) *The impact of social psychology on procedural justice*. Springfield, Il.: C.C. Thomas.

Wells, G.L. (1988). *Eyewitness identification: A system handbook.* Toronto, Ontario: Carswell legal publications.

Wells, G.L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553-71.

Wells, G.L. & Lindsay, R.C.L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*(3), 776-84.

Wells, G.L. & Loftus, E.F. (1984). (Eds.). *Eyewitness Testimony: Psychological Perspectives*. Cambridge: Cambridge University Press.

Wells, G.L. & Luus, E.C.A. (1990a). The diagnosticity of a lineup should not be confused with the diagnostic value of nonlineup evidence. *Journal of Applied Psychology, 75*(5), 511-516.

Wells, G.L. & Luus, E.C.A. (1990b). Police lineups as experiments: social methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin, 16*(1), 106-117.

Wells, G.L. & Turtle, J.W. (1986). Eyewitness identification: the importance of lineup models. *Psychological Bulletin, 99*(3), 320-29.

Wells, G.L. & Turtle, J.W. (1987). Eyewitness testimony research: current knowledge and emergent controversies. *Canadian Journal of Behavioural Science, 19*(4), 363-88.

Wells, G.L., Leippe, M.R. & Ostrom, T.M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behaviour, 3*(4), 285-293.

Wells, G.L., Rydell, S.M. & Seelau, E.P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*(5), 835-44.

Wells, G.L., Seelau, E.P., Rydell, S.M. & Luus, C.A.E. (1994). Recommendations for properly conducted lineup identification tasks. In D. Ross, J.D. Read, & Toglia, M.P. (Eds). *Adult Eyewitness testimony: Current trends and developments*. Pp. 223-44. New York: Cambridge University Press.

Whipple, G.M. (1912). Psychology of testimony and report. *Psychological Bulletin, 9,* 264-269.

Williams, G. & Hammelmann, H.A. (1955). Identification Parades. *Criminal Law Review*, 479-491, 545-555.

Wittgenstein, L. (1950/1978). *Philosophical investigations.* Tr. G.E.M. Anscombe. Oxford: Basil Blackwell.

Wogalter, M.S. & Laughery, K.R. (1987). Face recognition: Effects of study to test maintenance and change of photographic mode and pose. *Applied Cognitive Psychology, 1*(4) 241-253.

Wogalter, M.S. & Marwitz, D.B. (1991). Face composite construction: In-view and from-memory quality and improvement with practice. *Ergonomics, 34*(4), 459-468.

Woocher, F.D. (1977). Did your eyes deceive you? Expert testimony on the unreliability of eyewitness identification. *Stanford Law Review, 29,* 969-1030.

Yarmey, A.D. (1979). *Eyewitness testimony.* New York: Free Press.

Yarmey, A.D. (1993). Commentary: On ageing witnesses and earwitnesses. In G.M. Davies & R.H. Logie (Eds). *Memory in everyday life*. Pp 408-15. Elsevier Science Publishers.

Yarmey, A.D. (1994). Earwitness evidence: Memory for the perpetrator's voice. In D. Ross, J.D. Read, & M.P. Toglia (Eds). *Adult eyewitness testimony: Current trends and developments*. New York: Cambridge University Press.

Yin, R.K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*,141-145.

Young, A.W. (1992). Face recognition impairments. *Philosophical Transactions of the Royal Society of London, B, 335*, 47-54.

Young, A.W. (1993). Recognising friends and acquaintances. In G.M. Davies & R.H. Logie (Eds). *Memory in everyday life*. Pp 325-50. Elsevier Science Publishers.

Young, A., Hellawell, D. & Hay, D.C. (1987). Configural information in face perception. *Perception, 16*, 747-59.

Young, A.W. & Bruce, V. (1991). Perceptual categories and the computation of "grandmother". *European Journal of Cognitive Psychology, 3*(1), 5-49.

Young, A.W. & Hay, Dennis C. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology, 76*(4), *495-523.*

Young, A.W., Hay, D.C., McWeeny, K.H., Flude, B.M. & Ellis, A.W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, 14, 737-46.

Young, M.P. & Yamane, S. (1992) Sparse population coding of faces in the inferotemporal cortex. *Science*, *256* (May 29), 1327-31.

Appendices

Note that appendices include only certain materials referred to in the body of the thesis, or used in the empirical work. Most of the empirical studies used multiple sequences and orders, and only examples are typically included, since reproduction of all materials would result in an excessively long document.

Appendix A: Form SAP 329.

SUID-AFRIKAANSE POLISIE

UITKENNINGSPARADEVORM

Artikel 37 (1) (b) Wet Nr 51 van 1977

SOUTH AFRICAN POLICE

IDENTIFICATION PARADE FORM

Section 37 (1) (b) Act No 51 of 1977

1. Lid in beheer van parade
   Member in charge of parade.....
2. Stasie
   Station..... MR
   CR ...../...../.....
3. Klagte(s) • Charge(s) .....
   .....
   .....
   .....
4. Opdrag vir die hou van die parade ontvang op
   Instruction for the institution of the parade received on.....
   om
   at..... vanaf die ondersoekbeampte Nr
   from the investigating officer No.....
   Rang
   Rank..... Naam
   Name.....
5. Volle naam(e) van verdagte(s):
   Full name(s) of suspect(s):
   (1)..... Taal
   Language .....
   (2)..... Taal
   Language .....
   (3)..... Taal
   Language .....
   (4)..... Taal
   Language .....
   (In geval van meer verdagtes, vervolg op folio.) • (If there are more suspects, continue on folio.)
6. Verdagte(s) is op
   Suspect(s) was/were informed on ..... van die voorgenome
   of the intended
   parade op
   parade on ..... om
   at..... te
   at..... ingelig.
7. Verdagte(s) is op
   Suspect(s) was/were informed on..... verwittig dat hy/sy/hulle
   that he/she/they
   geregtig is op regsverteenwoordiging.
   is/are entitled to legal representation.
8. Verdagte(s) verlang/verlang nie regsverteenwoordiging/nie.
   Suspect(s) desire/do not desire legal representation.
9. Regsverteenwoordiger(s)
   Legal representative(s) (1)..... (2).....
   (3) ..... (4) ..... is
   was/were informed
   op
   on (1)..... (2) .....
   (3)..... (4).....
   ingelig van die datum, tyd en plek van die parade.
   of the date, time and place of the parade.
10. Naam van fotograaf
    Name of photographer .....
11. Naam van tolk
    Name of interpreter .....
12. Die parade is buite sig en gehoor van ander persone gehou.
    The parade was held out of sight and hearing of other persons.
    Plek
    Place ..... datum
    date ..... tyd
    time.....
13. Naam van lid wat toesig oor getuie(s) gehou het voordat hy/sy/hulle parade bygewoon het
    Name of member who guarded the witness(es) before he/she/they attended the parade .....
    Kantoor Nr
    .....Office No.....
14. Naam van lid wat getuie(s) na parade begelei
    Name of member who escorts witness(es) to the parade .....
    .....
15. Naam van lid wat getuie(s) van parade begelei
    Name of member who escorts witness(es) from the parade.....
    .....
16. Naam van lid wat toesig oor getuie(s) gehou het nadat hy/sy/hulle parade bygewoon het
    Name of member who guarded the witness(es) after he/she/they had attended the parade.....
    Kantoor Nr
    .....Office No.....
17. Daar was altesaam persone op parade [insluitende die verdagte(s)]. Hulle is min of meer van dieselfde lengte, liggaamsbou,
    There were..... .....persons on parade [including the suspect(s)]. They are of about the same height, build, age and appearance and
    ouderdom en voorkoms en is almal naastenby soos die verdagte(s) gekleed.
    were dressed more or less similar to the suspect(s).

18. Verdagte(s) is van die beweerde klagte(s) en die doel van die parade verwittig en meegedeel dat hy/sy/hulle—
Suspect(s) has/have been informed of the allegation(s) and the purpose of the parade and that he/she/they—
(1) enige posisie van sy/haar/hulle keuse op die parade mag inneem en van posisie mag verander voordat 'n ander getuie geroep word; en
may take up any position of his/her/their choice on the parade and may change his/her/their position before another witness is called; and
(2) enige redelike versoek(e) ten opsigte van die parade mag rig.
may make any reasonable request(s) in respect of the parade.

19. (1) Sy/haar/hulle versoek(e) is soos volg:
His/her/their request(s) is/are the following:

Verdagte • Suspect 1 .....

.....

Verdagte • Suspect 2 .....

.....

Verdagte • Suspect 3 .....

.....

Verdagte • Suspect 4 .....

.....

(2) Stappe gedoen as gevolg van die versoek(e):
Steps taken as a result of the request(s):

Verdagte • Suspect 1 .....

.....

Verdagte • Suspect 2 .....

.....

Verdagte • Suspect 3 .....

.....

Verdagte • Suspect 4 .....

.....

20. Verdagte(s) is gevra of hy/sy/hulle tevrede is met die opstel van die parade, insluitende die persone op parade.
Suspect(s) was/were asked whether he/she/they is/are satisfied with the parade, including the persons on parade.
Sy/haar/hulle antwoord(e) is soos volg:
His/her/their answer(s) is/are as follows:

Verdagte • Suspect 1 .....

.....

Verdagte • Suspect 2 .....

.....

Verdagte • Suspect 3 .....

.....

Verdagte • Suspect 4 .....

.....

21. Naam van regsverteenwoordiger(s), indien teenwoordig:
Name of legal representative(s), if present:

Naam • Name	Namens • On behalf of

22. Persone op parade, verdagte(s) ingesluit:
Persons on parade, suspect(s) included:

Naam • Name	Ouderdom • Age	Adres • Address
(1)		
(2)		
(3)		
(4)		
(5)		
(6)		
(7)		
(8)		
(9)		
(10)		
(11)		

Naam • Name	Ouderdom • Age	Adres • Address
(12)		
(13)		
(14)		
(15)		
(16)		
(17)		
(18)		
(19)		
(20)		

23. 'n Foto is geneem nadat die parade opgestel is: Ja • Nee
A photograph was taken after the parade had been set up: Yes • No

VERLOOP VAN PARADE • PROCEDURE OF PARADE

24. Die persone, insluitende die verdagte(s), wat aan die parade deelneem, neem die volgende posisies van links na regs voor my in:
The persons, including the suspect(s), taking part in the parade, take up the following positions from left to right in front of me:

(1)..... (2).....

(3)..... (4).....

(5)..... (6).....

(7)..... (8).....

(9)..... (10).....

(11)..... (12).....

(13)..... (14).....

(15)..... (16).....

(17)..... (18).....

(19)..... (20).....

25. Eerste getuie • First witness

Naam • Name..... Taal • Language .....
is gevra om die verdagte(s), indien op parade, uit te wys deur sy/haar/hul skouer(s) aan te raak, wat (datum, tyd, plek en klagte):
was asked to point out the suspect(s), if on parade, by touching his/her/their shoulder(s), who (date, time, place and charge):

.....

.....

.....

(1) Tyd deur getuie geneem om persoon(e) op parade uit te wys
Time taken by witness to point out person(s) on parade.....

Uitsleg • Result .....

.....

(2) Reaksie van getuie tydens uitwysing van persoon(e):
Reaction of witness during pointing out of person(s):

.....

.....

26. Verdagte(s) word geleentheid gebied om van posisie(s) te verander en gevra of hy/sy/hulle tevrede is met sy/haar/hulle posisie. Sy/haar/hulle Suspect(s) is/are given the opportunity of changing his/her/their position(s) and asked whether he/she/they is/are satisfied. His/her/their antwoord(e) was:
answer(s) was/were: (1).....

(2).....

(3).....

(4).....

27. Posisies deur persone op parade ingeneem voordat tweede getuie op parade verskyn:
Positions taken by persons on parade before the second witness appears on parade:

(1)..... (2).....

(3)..... (4).....

(5)..... (6).....

(7)..... (8).....

(9)..... (10).....

(11)..... (12).....

(13)..... (14).....

(15)..... (16).....

(17)..... (18).....

(19)..... (20).....

28. Tweede getuie • Second witness:

Naam • Name..... Taal • Language .....
is gevra om die verdagte(s), indien op parade, uit te wys deur sy/haar/hul skouer(s) aan te raak, wat (datum, tyd, plek en klagte):
was asked to point out the suspect(s), if on parade, by touching his/her/their shoulder(s), who (date, time, place and charge):

.....

.....

.....

(1) Tyd deur getuie geneem om persoon(e) op parade uit te wys
Time taken by witness to point out person(s) on parade.....

Uitslag • Result .....

.....

(2) Reaksie van getuie tydens uitwysing van persoon(e):
Reaction of witness during pointing out of person(s):

.....

.....

29. Verdagte(s) word geleentheid gebied om posisie(s) te verander en gevra of hy/sy/hulle tevrede is met sy/haar/hulle posisie. Sy/haar/hulle Suspect(s) is/are given the opportunity of changing his/her/their position(s) and asked whether he/she/they is/are satisfied. His/her/their antwoord(e) was:
answer(s) was/were:

(1).....

(2).....

(3).....

(4).....

30. Posisies deur persone op parade ingeneem voordat derde getuie op parade verskyn:
Positions taken by persons on parade before the third witness appears on parade:

(1)..... (2).....

(3)..... (4).....

(5)..... (6).....

(7)..... (8).....

(9)..... (10).....

(11)..... (12).....

(13)..... (14).....

(15)..... (16).....

(17)..... (18).....

(19)..... (20).....

31. Derde getuie • Third witness:

Naam • Name..... Taal • Language .....
is gevra om die verdagte(s), indien op parade, uit te wys deur sy/haar/hul skouer(s) aan te raak, wat (datum, tyd, plek en klagte):
was asked to point out the suspect(s), if on parade, by touching his/her/their shoulder(s), who (date, time, place and charge):

.....

.....

.....

(1) Tyd deur getuie geneem om persoon(e) op parade uit te wys
Time taken by witness to point out person(s) on parade.....

Uitslag • Result .....

.....

(2) Reaksie van getuie tydens uitwysing van persoon(e):
Reaction of witness during pointing out of person(s):

.....

.....

32. Opmerkings, indien enige
Remarks, if any .....

.....

.....

.....

33. Naam van polisiestasie
Name of police station .....

VB Nr
OB No. ...../...../.....

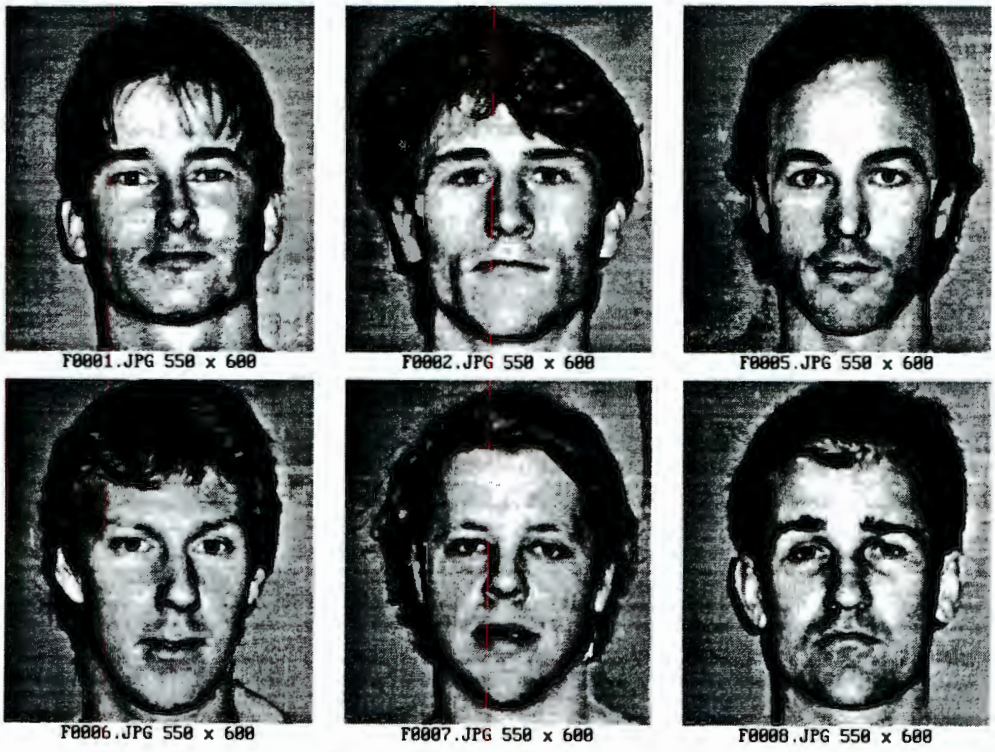
34. Ek, Nr Rang Naam
I, No. ..... Rank ..... Name.....

sertifiseer dat hierdie parade deur my waargeneem is, dat die besonderhede wat op die vorm deur my ingevul, korrek en 'n presiese weergawe van die gebeure is.
certify that this parade was conducted by me, that the particulars which have been completed on the form by me are correct and that it is a just report of the procedures which took place.

.....
Handtekening, Nr en rang van lid in beheer van parade
Signature, No and rank of member in charge of parade

Appendix B: Arrays of faces used in experiments 1a and 1b

Faces used in experiment 1a



Faces used in experiment 1b

Appendix C: The collection of images used in Study 2

Appendix D : Frontal views collected for Studies 3-7

Note: Faces shown here are not in final standardized form.

Appendix E : Sample experimental booklet used in Study 3

Face Recognition Study

Colin Tredoux
Department of Psychology
University of Cape Town
Rondebosch 7700

Over the page you will find three collections of faces. The collections are marked "A", "B", and "C", respectively.

In each collection you will notice that each face has a number below it, except one, which is called the 'target' face. I would like you to compare the numbered faces to the target face - how similar in facial appearance are they to the target?

In order to arrive at the rating of similarity, use any facial quality you think relevant. You may also wish to take the following into consideration:

Hair: are the length, colour, and texture similar?
Hairline: is the hairline equally high or low on the forehead?
Face shape: do the faces have the same shape (e.g. round, thin, angular)?
Noses: are the noses similar in size and shape?
Mouths: are the mouths equal in size? are the lips equally full?
Skin texture and colour: is this similar?
Chins: are these the same shape and size?

Please use a scale with the extreme values shown below. You can assign any number between 0 and 10.

0_____ 5_____ 10

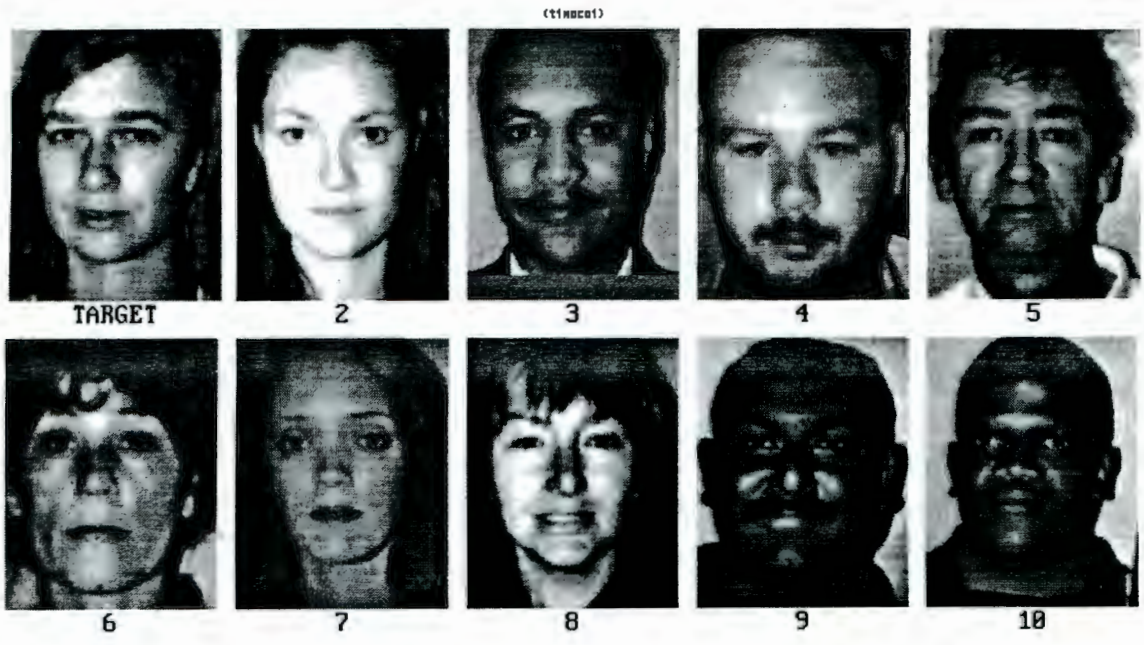
Not at all similar Highly similar

Please indicate for each face, how similar it is to the target face. Do this by writing the value you have chosen next to the number of the face. Repeat this process for each of the three collections, "A", "B" and "C".

Thank you for participating in this study.

A

B



C

Appendix F : Profile views (of face images) used in Study 4

Note: images are not in final standardized form.

Appendix G : Sample experimental booklet used in Study 4

Face Recognition Study

Colin Tredoux
Department of Psychology
University of Cape Town
Rondebosch 7700

Over the page you will find two collections of faces. The collections are marked "A", and "B", respectively.

In each collection you will notice that each face (or pair of faces) has a number below it, except one, which is called the 'target' face. I would like you to compare the numbered faces to the target face - how similar in facial appearance are they to the target?

In order to arrive at the rating of similarity, use any facial quality you think relevant. You may also wish to take the following into consideration:

Hair: are the length, colour, and texture similar?
Hairline: is the hairline equally high or low on the forehead?
Face shape: do the faces have the same shape (e.g. round, thin, angular)?
Noses: are the noses similar in size and shape?
Mouths: are the mouths equal in size? are the lips equally full?
Skin texture and colour: is this similar?
Chins: are these the same shape and size?

Please use a scale with the extreme values shown below. You can assign any number between 0 and 10.

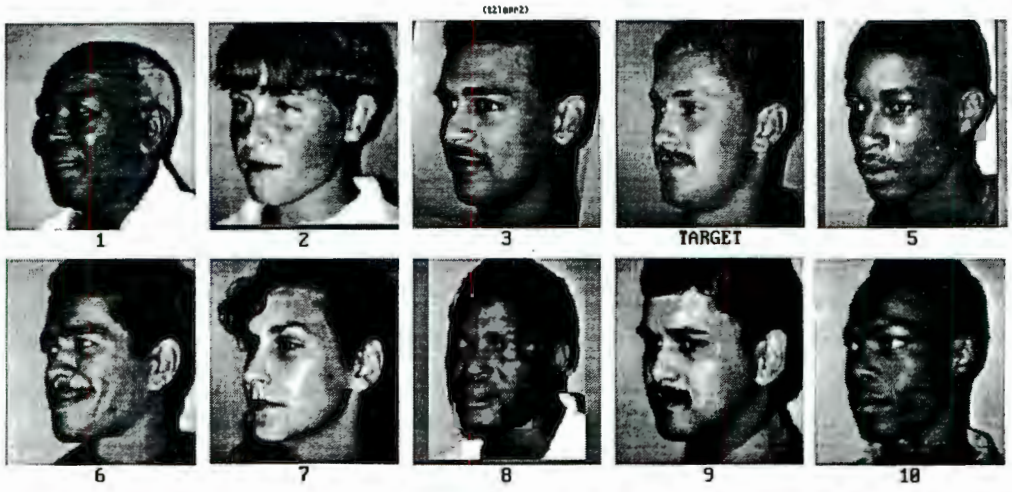
0_____ 5_____ 10

Not at all similar Highly similar

Please indicate for each face, how similar it is to the target face. Do this by writing the value you have chosen next to the number of the face. Repeat this process for each of the collections, "A" and "B".

Thank you for participating in this study.

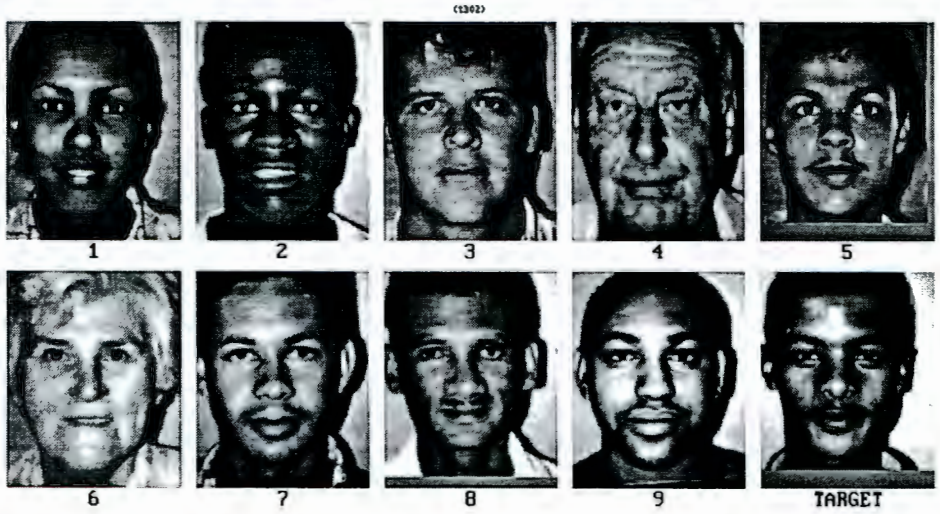

1
TARGET
3
4
5
6
7
8
9
10


(121BMP2)
1
2
3
TARGET
5
6
7
8
9
10

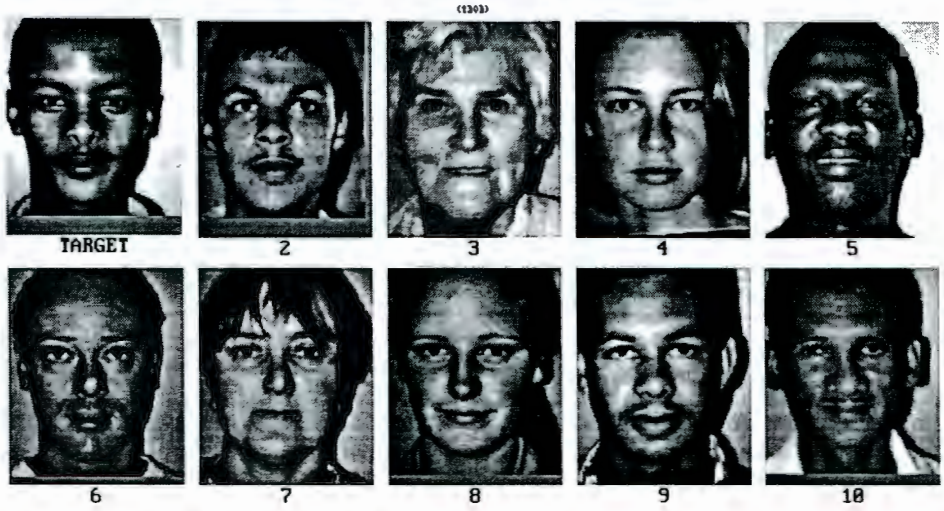
Appendix H : Arrays used in Study 5

Initial array

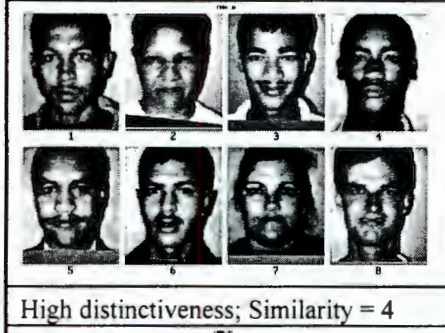
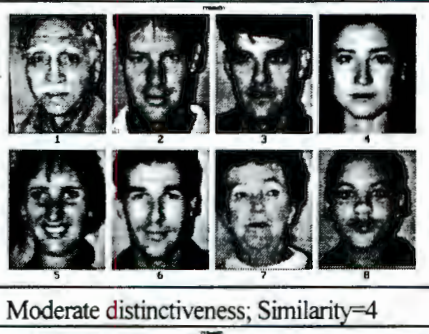
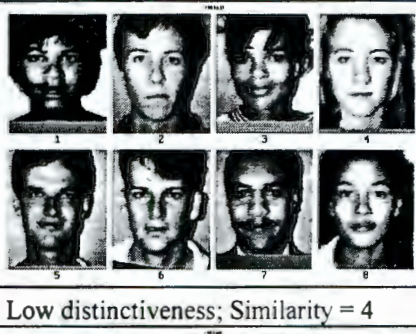
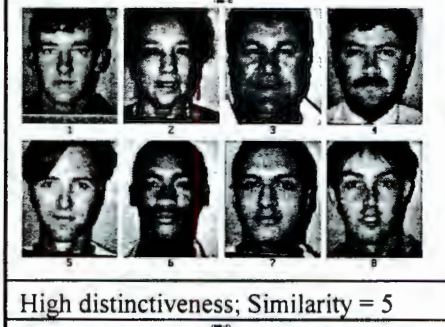
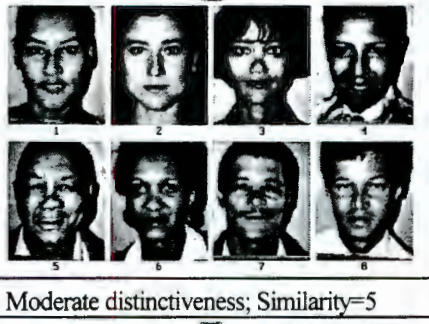
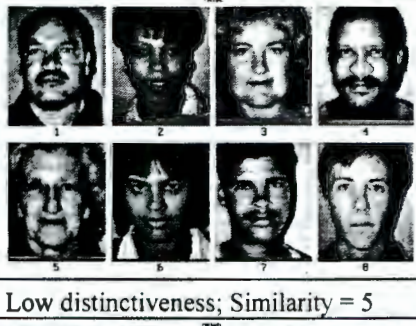
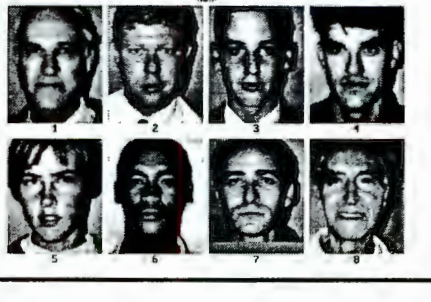
Follow-up array 1



Follow-up array 2



Appendix I : Parades used in the mock witness task of Study 6

High distinctiveness; Similarity = 1	Moderate distinctiveness; Similarity=1	Low distinctiveness; Similarity = 1
		
High distinctiveness; Similarity = 2	**Moderate distinctiveness; Similarity=2**	**Low distinctiveness; Similarity = 2**
		
High distinctiveness; Similarity = 3	**Moderate distinctiveness; Similarity=3**	**Low distinctiveness; Similarity = 3**
		
High distinctiveness; Similarity = 4	**Moderate distinctiveness; Similarity=4**	**Low distinctiveness; Similarity = 4**
		
High distinctiveness; Similarity = 5	**Moderate distinctiveness; Similarity=5**	**Low distinctiveness; Similarity = 5**
		

High distinctiveness; Similarity = 6	Moderate distinctiveness; Similarity=6	Low distinctiveness; Similarity = 6
		

Appendix J : Documents given to subjects at the initial stage of the experiment in Study 7

Document A

Below, you will find pictures and descriptions of three students. The details were taken from a self-report questionnaire completed by the students. Please read the descriptions, and look at the pictures.

Cassiem

Age 26. Majors are English and History. Would like to become a teacher after completing first degree. Is interested in movies, dancing and music.

Susan

Age 37. Busy completing an honours degree in Archaelogy. Married, with three children. Is interested in walking, bird-watching, and heraldry.

Astrid

Age 20. First-year B.Sc. student. Not sure of a career track. Is interested in volleyball, walking on the beach, movies, and likes working with children.

Please turn the page.

Now, try to think of one further thing which you think is probably true of each of the three people on the previous page (without turning back to the descriptions)

Cassiem:

1 _____

Susan:

1 _____

Astrid:

1 _____

Document B

Comparative Perception Study

Colin Tredoux
Department of Psychology
University of Cape Town
Rondebosch 7700

Below, you will find pictures and descriptions of three students. The details were taken from a self-report questionnaire completed by the students. Please read the descriptions, and look at the pictures.

David

Age 24. Majors are English and History. Would like to become a teacher after completing first degree. Is interested in movies, dancing and music.

Flip

Age 25. Busy completing an honours degree in Archaelogy. Married, with three children. Is interested in walking, bird-watching, and heraldry.

Bongani

Age 30. First-year Soc.Sci. student. Not sure of a career track. Is interested in politics, ball-room dancing, and black-and-white film technology.

Please turn the page.

Now, try to think of one further thing which you think is probably true of each of the three people on the previous page (without turning back to the descriptions)

David:

1 _____

Flip:

1 _____

Bongani:

1 _____

Appendix K : Test instructions, and lineups, used in the test phase of the experiment in Study 7

1 Test instructions for subjects who completed document 'A' in the initial phase

Comparative Perception Study

Colin Tredoux
Department of Psychology
University of Cape Town
Rondebosch 7700

Earlier, you were provided with pictures and descriptions of three students. What I would like you to do now is to point them out - if they appear - in the collections of pictures that follow.

1 Over the page you will find the first collection of numbered pictures (I have marked it as 'A'). If the student *Cassiem* appears in this collection, indicate which number is below his picture (if he is not present, mark the checkbox):

Cassiem does not appear ☐ *Cassiem* is number ___

2 Turn the page to the next collection of numbered pictures (I have marked it as 'B'). If the student *Susan* appears in this collection, indicate which number is below her picture (if she is not present, mark the checkbox):

Susan does not appear ☐ *Susan* is number _____

3 Turn the page to the next collection of numbered pictures (I have marked it as 'C'). If the student *Astrid* appears in this collection, indicate which number is below her picture (if she is not present, mark the checkbox):

Astrid does not appear ☐ *Astrid* is number _____

Thank you for participating in this study

2 Test instructions for subjects who completed document 'B' in the initial phase

Comparative Perception Study

Colin Tredoux
Department of Psychology
University of Cape Town
Rondebosch 7700

Earlier, you were provided with pictures and descriptions of three students. What I would like you to do now is to point them out - if they appear - in the collections of pictures that follow.

1 Over the page you will find the first collection of numbered pictures (I have marked it as 'A'). If the student *David* appears in this collection, indicate which number is below his picture (if he is not present, mark the checkbox):

David does not appear ☐ *David* is number _____

2 Turn the page to the next collection of numbered pictures (I have marked it as 'B'). If the student *Flip* appears in this collection, indicate which number is below her picture (if he is not present, mark the checkbox):

Flip does not appear ☐ *Flip* is number _____

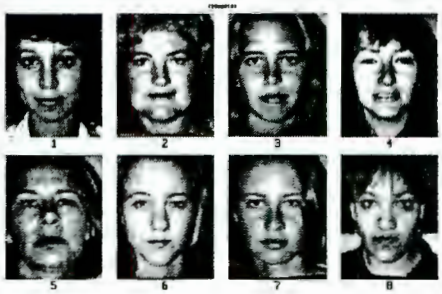
3 Turn the page to the next collection of numbered pictures (I have marked it as 'C'). If the student *Bongani* appears in this collection, indicate which number is below her picture (if he is not present, mark the checkbox):

Bongani does not appear ☐ *Bongani* is number ____

Thank you for participating in this study

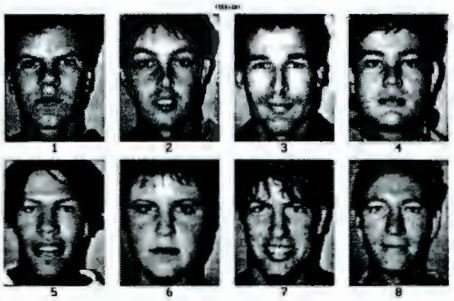
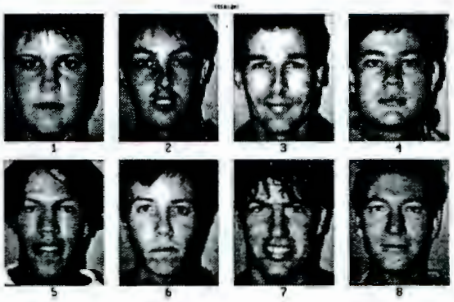
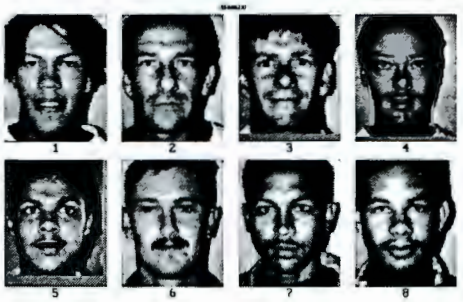
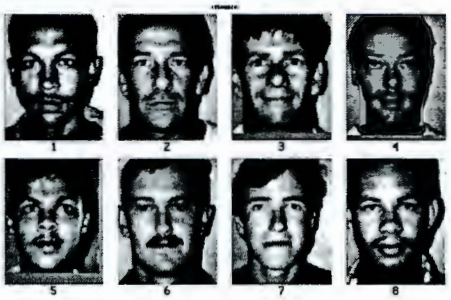
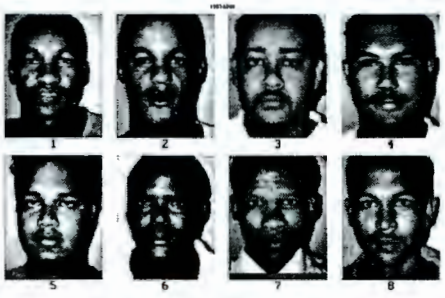
3 Lineups used in tests of subjects who completed document 'A' in the initial phase

(The same set of images was used for sequential lineups).

SIMILARITY	**Target absent**	**Target present**
High (Target is no 3)		
Moderate (Target is no 4)		
Low (Target is no 5)		

4 Lineups used in tests of subjects who completed document 'B' in the initial phase

(The same set of images was used for sequential lineups).

SIMILARITY	Target absent	Target present
High (Target is no 6)		
Moderate (Target is no 7)		
Low (Target is no 6)		

Appendix L : Raw data for Studies 1 - 7

Note that in cases where different sequences and orders were used, data are reported according to the original sequence/order, and are not comparable across sequence/order.

Study 1b

			Array member				
Subject	Order	Sequence	A	B	C	D	E
1	1	1	12	11	8	10	9
2	1	1	9	8	10	11	12
3	2	1	12	11	10	8	9
4	1	1	8	10	11	12	9
5	2	1	8	12	10	11	9
6	1	1	12	10	8	4	11
7	2	1	9	8	10	11	12
8	2	1	8	11	10	9	12
9	1	1	12	10	8	11	9
10	2	1	11	12	8	9	10
11	1	1	12	11	8	10	9
12	1	1	12	9	8	10	11
13	2	2	8	9	10	12	11
14	1	2	8	9	11	12	10
15	2	2	8	10	12	11	9
16	2	2	10	8	11	9	12
17	1	2	12	11	9	10	8
18	1	2	8	9	11	12	10
19	1	2	10	12	8	11	9
20	2	2	8	10	11	9	12
21	2	2	10	8	12	9	11
22	1	2	10	8	11	9	12
23	1	2	8	12	11	9	10
24	1	3	10	12	8	9	11
25	1	3	11	12	10	8	9
26	2	3	12	11	9	8	10
27	1	3	8	9	12	10	11
28	2	3	12	10	8	9	11
29	2	3	8	10	9	12	11
30	2	3	8	12	10	11	9
31	1	3	12	8	10	11	9
32	1	3	12	8	10	11	9
33	2	3	12	8	11	9	10
34	2	3	9	10	8	11	12

Study 1a

			Array member				
Subject	Order	Sequence	A	B	C	D	E
1	2	1	4	5	3	2	1
2	2	1	1	5	4	2	3
3	2	1	1	5	3	2	4
4	2	1	1	5	4	3	2
5	2	1	1	2	5	4	3
6	1	1	1	4	3	2	5
7	2	1	1	2	4	3	5
8	2	1	1	5	2	3	4
9	2	1	1	2	5	4	3
10	2	1	1	5	4	3	2
11	1	1	2	3	4	1	5
12	2	1	1	4	2	3	5
13	1	1	1	4	3	2	5
14	1	1	1	4	5	3	2
15	2	1	1	2	5	3	4
16	1	1	1	2	4	5	3
17	2	2	3	5	4	2	1
18	1	2	2	1	5	4	3
19	1	2	2	4	5	3	1
20	1	2	1	2	5	3	5
21	1	2	3	4	5	2	1
22	1	2	1	5	4	2	3
23	1	2	1	2	5	4	3
24	2	2	1	2	3	5	4
25	1	2	2	4	5	3	1
26	2	2	2	1	4	3	5
27	1	2	1	5	4	2	3
28	2	2	1	3	4	5	2
29	1	3	1	3	5	4	2
30	2	3	1	2	4	3	5
31	2	3	1	2	3	5	4
32	1	3	1	2	5	4	3
33	2	3	2	1	5	4	3
34	1	3	2	4	5	1	3
35	2	3	1	2	4	3	5
36	1	2	1	2	6	3	4

Data are ranks. In Study 1a, the ranks progress from 1-5, and in Study 1b, from 8-12. Thus, an '8' for member A in Study 1b means that the corresponding face image is ranked most similar to the target.

Study 2 : Rated distinctiveness data

Face images were rated on a 15 point scale. The numbers in the first row (below 'Array a') follow the sequence of images in arrays presented to subjects. Arrays 4-6 were different sequences of images used in 1-3.

Array 1										Array 2										Array 3									
1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
1	1	1	7	15	1	1	10	1	1	1	1	1	10	1	1	10	15	1	1	1	1	1	15	7	7	1	15	1	1
3	3	3	7	4	3	7	7	4	3	3	3	3	14	7	3	7	5	3	4	3	3	3	7	3	3	3	7	3	3
1	1	7	7	7	7	1	15	1	1	7	7	7	15	1	7	15	7	1	1	1	1	1	15	1	1	1	7	1	1
15	1	7	7	15	1	1	7	1	7	1	1	1	15	15	7	1	1	7	1	7	1	1	15	15	1	1	1	15	15
8	4	9	2	3	2	2	3	4	7	6	7	3	8	3	9	3	3	7	4	3	7	4	8	5	3	3	4	8	8
15	7	1	15	15	15	7	1	1	7	15	7	7	1	15	1	15	7	7	7	1	7	15	1	7	15	7	1	1	15
7	7	7	15	1	15	7	7	7	15	7	7	7	15	7	1	7	7	15	7	7	15	7	7	1	1	1	7	7	7
1	1	7	1	1	1	1	1	1	1	1	1	1	7	1	7	1	1	1	1	1	1	1	7	1	1	1	1	1	7
14	13	10	13	14	12	13	14	9	4	8	5	6	8	11	10	14	13	4	7	9	8	8	10	11	9	10	13	12	12
1	1	15	7	15	1	7	15	1	1	7	15	7	7	7	15	15	15	1	1	7	1	1	15	7	1	15	1	15	7
15	15	7	15	7	15	7	1	15	15	15	15	15	7	7	15	7	7	15	15	15	15	7	7	15	15	15	15	15	15
15	8	14	14	13	4	4	5	5	2	13	10	2	8	15	14	4	7	2	2	4	10	8	14	14	8	5	10	7	9
4	4	8	3	1	7	6	14	3	5	1	13	2	9	12	7	14	7	6	6	15	5	14	13	9	15	7	1	1	6
1	1	6	1	8	1	1	2	1	1	1	1	1	1	3	6	2	8	1	1	1	3	3	1	1	1	1	1	1	1
11	10	8	7	1	8	12	2	8	7	7	8	8	9	4	8	4	11	7	6	7	7	13	13	7	9	3	9	4	10

Array 4										Array 5										Array 6									
13	7	7	6	6	6	5	13	7	6	14	13	7	5	6	4	8	7	7	7	7	6	10	5	5	5	7	7	7	8
1	15	7	1	15	7	1	15	15	7	15	15	1	1	1	1	15	7	7	7	7	7	7	7	1	1	7	15	7	1
7	1	1	1	1	1	1	1	7	7	15	15	1	1	7	1	7	7	1	1										
4	14	3	5	6	7	5	9	10	11	2	8	6	3	5	3	2	7	3	2	2	3	5	4	2	5	3	8	3	7
7	13	5	10	7	5	3	15	10	4	15	10	3	5	2	2	3	3	10	5	3	6	5	5	6	4	3	11	14	11
7	13	7	10	13	10	3	10	3	10	15	10	6	3	7	7	3	13	10	7	13	15	7	7	4	4	7	14	13	13
7	15	7	7	7	7	7	15	15	7	15	15	7	7	7	4	7	4	6	7	3	3	5	3	3	5	5	7	7	7
7	9	8	7	7	7	6	15	6	5	15	10	8	7	10	6	8	5	8	6	4	5	15	6	6	6	8	8	8	7
7	1	15	15	15	7	7	7	7	15	7	7	15	7	15	7	7	7	7	7	7	7	7	7	7	15	7	1	15	7
7	15	1	7	7	1	7	15	1	1	15	15	7	1	1	7	1	1	7	1	1	15	15	1	7	1	15	15	1	7
7	2	7	15	15	7	2	1	7	7	1	1	7	1	15	7	15	7	7	15	7	1	1	7	15	7	7	1	15	7
7	15	8	1	12	9	1	15	15	7	14	12	7	9	2	1	3	7	6	1	5	13	14	7	3	6	3	1	1	1
7	12	7	8	8	6	6	8	13	12	12	13	6	5	6	5	6	5		6	5	10	12	4	6	4	4	10	11	8
7	12	4	4	6	6	2	7	4	9	13	10	5	3	3	5	8	3		3	5	7	11	3	3	7	3	12	8	4
10	13	10	7	7	7	7	9	7	9	13	12	6	7	5	6	10	6	6	6	5	6	5	5	5	5	6	10	7	7
7	15	7	7	1	1	1	15	1	1	15	15	1	15	7	1	15	7	7	7	7	7	15	1	1	1	7	15	7	7
15	15	7	1	15	15	1	1	7	1	1	15	15	7	15	15	1	7	7	1	15	1	1	7	7	15	15	15	7	7

Study 2: Rated typicality data

Face images were rated on a 15 point scale. The numbers in the first row (below 'Array a') follow the sequence of images in arrays presented to subjects.

	Array 1										Array 2										Array 3									
Sequence	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
a	2	7	13	1	1	3	3	7	14	3	15	5	1	7	7	12	6	3	1	9	2	15	9	6	1	10	7	4	7	15
a	7	7	1	15	7	1	1	1	15	15	7	7	1	1	7	1	15	1	1	7	1	1	1	15	1	7	7	1	7	15
a	15	10	12	11	13	12	15	4	8	10	14	10	3	9	10	12	8	12	12	15	4	8	8	5	1	2	7	12	8	10
a	14	7	12	10	9	10	6	7	15	10	15	7	3	7	15	10	9	7	12	14	3	5	10	6	1	14	12	6	10	15
a	15	7	7	1	7	1	7	1	15	15	7	7	1	1	7	1	15	7	1	15	1	7	1	7	7	15	15	7	15	15
a	15	15	15	15	15	15	15	7	7	15	15	1	7	1	15	1	1	15	1	15	15	1	15	15	1	15	1	15		15
a	15	15	7	7	7	7	1	1	7	15	15	7	15	1	7	15	15	7	1	7	15	7	7	1	1	1	7	1	1	15
a	10	10	10	1	8	5	6	2	6	12	8	10	8	2	2	7	4	4	8	10	2	15	5	6	6	10	10	6	8	6
a	7	7	15	7	7	15	1	1	15	1	7	15	1	1	7	15	15	1	15	7	1	15	15	7	7	7	15	7	1	7
a	7	15	7	7	13	15	7	15	13	7	15	1	7	1	7	7	1	7	7	7	1	15	15	1	1	15	7	7	15	7
a	9	7	7	5	2	6	9	5	12	13	12	10	5	5	14	14	4	5	8	14	9	9	6	2	1	13	12	9	10	12
a	7	15	1	7	1	12	7	15	15	7	7	3	1	2	7	12	7	12	2	15	1	5	12	7	7	12	1	7	1	15
a	1	1	7	15	7	1	7	7	15	7	7	15	15	7	1	15	15	1	7	15	15	1	1	1	1	1	1	1	1	15
a	8	6	13	9	7	6	3	1	10	7	5	7	4	4	7	7	8	5	5	10	4	5	6	7	5	6	5	3	3	10
a	12	9	11	6	5	4	1	3	11	3	9	7	3	7	15	7	7	6	11	15	1	12	4	7	2	4	2	1	7	12
a	2	10	11	7	3	10	4	1	9	6	3	5	4	8	2	4	7	6	9	7	4	10	10	3	4	8	8	5	7	11
a	12	13	8	13	8	13	6	13	13	12	14	13	12	11	12	13	13	13	13	13	8	12	13	3	3	10	12	3	11	10
a	7	7	15	7	15	7	15	7	7	7	7	15	15	7	7	7	7	1	7	7	1	15	7	7	7	7	7	15	7	7
a	7	1	7	7	7	15	7	1	15	7	1	1	15	7	15	1	15	7	7	7	1	7	15	1	7	7	15	7	7	15
b	10	5	11	4	12	12	13	7	5	6	7	7	7	8	8	14	14	7	10	3	6	9	12	9	3	7	10	3	10	13
b	7	15	1	15	1	1	1	1	1	1	1	1	1	1	1	1	15	1	1	7	1	1	1	7	1	1	1	1	1	1
b	10	10	7	1	7	10	15	1	1	7	15	7	10	7	10	7	13	15	7	7	7	7	5	10	7	7	13	5	5	10
b	7	15	7	1	7	1	1	7	1	7	15	7	7	15	7	15	7	15	7	15	15	15	15	7	1	15	15	15	15	7
b	5	5	4	10	1	1	1	7	2	2	6	5	8	11	6	3	5	2	6	5	5	9	11	5	2	7	1	3	6	6
b	12	1	7	1	1	1	3	1	1	1	6	1	1	1	6	12	1	1	1	1	1	1	12	12	7	1	1	1	1	1
b	10	12	7	5	9	7	12	9	5	4	4	9	4	8	7	7	6	3	4	4	5	7	12	10	8	5	7	5	6	8
b	12	12	7		12	7	7	5	8	10	10	10	6	6	8	12	13	6	8	6	9	10	12	12	7	7	7	5	9	12
b	7	8	10	9	13	13	6	13	13	13	13	13	12	13	13	13	13	6	13	11	9	13	13	8	5	13	13	13	13	13
b	8	6	8	12	10	8	6	5	12	10	8	10	4	6	9	12	12	4	10	6	4	10	12	8	15	10	10	5	10	10
b	15	15	15	15	15	15	15	7	7	15	15	15	1	7	15	7	13	7	15	7	7	15	15	7	1	15	15	7	15	15
b	7	10	3	8	8	5	4		3	8	9	5	3	8	8	5	15	3	8	15	2	5	4	7	1	7	4	3	9	15
b	7	15	15	1	1	1	1	9	1	15	10	13	3	7	10	7	15	2	9	15	2	8	13	7	1	15	9	1	4	15
b	7	7	10	3	7	10	10	3	5	5	5	10	3	7	12	10	3	7	10	3	5	7	10	7	5	8	8	3	5	8
b	15	15	15	15	15	7	15	1	1	7	15	15	10	15	15	15	15	4	15	12	3	10	15	7	3	14	14	2	13	13
b	1	1	1	1	1	1	1	1	1	1	1	1	1	10	1	1	15	1	1	1	1	1	1	1	1	1	1	1	1	1
b	7	5	4	4	6	8	4	4	7	4	7	7	2	5	7	4	8	1	6	6	3	7	7	6	1	4	5	5	10	12
b	7	7	1	15	1	7	7	7	7	1	7	7	15	1	15	1	15	1	15	7	7	15	7	7	1	1	1	7	1	7

Study 2: Ranked similarity task

Numbers in the first row follow the sequence of the original arrays: missing numbers indicate the presence of the target in the corresponding position.

	Array white male										Array white female									Array black male									
Sequence	1	2	4	5	6	7	8	9	10		2	3	4	5	6	7	8	9	10		1	2	3	4	5	7	8	9	10
a	9	5	6	10	3	4	7	2	8		1	4	9	6	7	8	5	3	2		2	4	7	1	8	9	3	5	6
a	9	3	4	6	2	8	5	1	7		1	4	7	3	8	9	5	6	2		1	4	2	3	5	8	9	7	6
a	8	3	7	6	4	9	1	5	2		2	5	6	9	8	7	1	4	3		10	2	6	3	7	9	8	5	4
a	8	3	5	6	2	4	7	4	1		1	4	5	2	7	8	9	3	6		4	1	2	3	8	9	7	5	6
a	8	4	1	3	6	5	2	7	9		1	4	5	7	9	8	3	2	6		1	7	4	2	8	6	5	3	9
a	8	7	5	9	6	4	1	2	3		1	2	6	5	9	7	8	3	4		6	2	5	1	8	3	7	4	9
a	10	6	3	9	5	8	4	2	7		2	7	10	6	9	8	5	3	4		2	6	9	8	3	5	7	10	4
a	6	4	8	9	7	2	5	3	1		1	3	8	9	7	4	2	6	5		5	7	1	3	9	8	6	2	4
a	6	3	8	9	5	7	1	2	4		1	7	9	5	4	8	3	2	6		1	6	5	2	9	5	7	8	4
a	7	6	5	9	8	3	1	4	2		2	1	9	7	8	6	3	5	4		10	2	6	5	9	3	8	7	4
a	7	2	8	9	1	3	5	4	6		1	2	4	6	8	7	9	5	3		2	7	4	8	1	5	9	6	3
a	7	5	3	8	9	2	1	6	4		2	5	7	6	9	8	4	1	3		1	4	2	3	6	5	9	7	8
a	9	7	6	8	3	4	5	2	1		1	5	3	8	9	6	7	2	4		8	3	2	1	9	4	5	6	7
a	8	2	7	9	3	1	4	5	6		1	7	6	4	8	9	5	3	2		4	3	6	1	5	8	9	2	7
a	5	2	7	8	6	3	9	4	1		1	7	9	2	8	6	5	3	4		2	5	4	1	7	9	8	3	6
a	9	7	2	8	4	6	5	3	1		1	5	6	4	8	9	7	3	2		9	6	2	4	10	7	8	3	5
b	3	7	5	8	2	9	3	4	1		2	3	10	4	7	9	8	6	5		1	4	5	3	8	9	10	7	6
b	4	8	6	9	1	3	2	5	7		2	1	7	8	6	9	3	4	5		2	1	6	3	7	8	9	5	4
b	9	6	3	10	5	1	2	4	8		3	2	5	6	4	9	7	8	1		1	4	2	3	5	6	9	8	7
b	9	2	1	4	5	3	6	8	7		2	8	5	6	1	9	7	3	4		1	4	6	3	7	9	8	2	5
b	3	6	5	10	8	7	4	2	9		1	5	7	4	8	9	2	3	6		1	6	7	3	5	4	8	9	2
b	7	8	6	9	3	4	1	2	5		1	4	6	7	8	5	9	2	3		7	2	4	1	8	9	5	3	6
b	8	6	2	7	5	4	3	1	9		1	5	8	7	9	6	3	2	4		1	5	4	2	6	8	9	7	3
b	7	5	2	4	3	8	1	6	9		2	6	7	5	8	9	3	1	4		1	2	6	5	3	4	8	7	9
b	9	4	1	8	3	6	5	2	7		4	5	1	8	9	6	7	2	3		2	3	4	6	8	9	7	5	1
b	7	1	2	3	5	6	8	4	9		1	4	9	5	7	8	6	2	3		4	2	3	1	7	5	8	6	9
b	9	7	1	8	6	5	3	4	2		1	3	8	7	2	9	5	6	4		2	8	5	1	3	9	7	4	6
b	9	6	3	10	5	1	2	4	8		3	2	5	6	4	9	7	8	1		1	4	2	3	5	6	9	8	7

Study 2: Similarity pairing task

Numbers in the first row indicate pairings. Numbers in the second row indicate the identity of the images making up the pairs (i.e. '6' and '20' means that subject 1 rated the 6th and 20th images in the array as most similar).

Array no																						
a	6	20	8	15	1	18	9	19	4	14	5	11	7	12	2	16						
a	9	7	3	6	2	5	1	4	8	10	14	19	11	12	16	17	18	20	20	15	13	8
a	16	17	15	20	14	19	12	15	11	20	1	8	6	9	10	3	2	4	6	3	7	10
a	16	17	1	5	3	6	15	20	14	15	11	12	10	4	8	9	7	2	13	18		
a	8	9	3	6	7	10	16	17	11	13	4	19	14	15	18	20	2	5	4	14	6	9
a	6	3	4	10	1	8	5	9	7	2	18	20	14	19	13	17	11	12	16	17	18	15
a	2	3	7	10	8	9	4	5	1	6	14	15	18	20	11	13	12	19	16	17		
a	4	14	18	20	3	10	16	17	6	2	7	15	8	13	5	11	9	19	1	12		
a	14	19	3	18	1	15	5	11	2	17	9	13	10	20	4	16	6	7	8	17		
a	3	10	14	19	16	17	5	11	1	18	8	13	2	6								
a	1	18	2	13	3	19	4	11	5	12	6	17	7	20	8	14	9	16	10	15		
a	3	18	1	20	2	15	4	19	5	11	6	14	7	13	8	17	9	12	10	16		
a	3	6	2	10	7	20	9	15	14	18	5	13	1	4	8	19	16	17	11	12		
a	11	12	14	19	2	9	16	17	7	10	3	6	1	8	15	20	4	5	16	18		
a	7	10	3	6	19	20	2	15	3	14	8	9	11	12								
a	14	19	8	9	1	5	15	10	11	12	2	4	3	20								
a	3	14	6	10	1	9	7	20	2	19	5	17	8	13	4	18	11	12	14	16	1	15
a	4	14	7	20	8	18	9	19	10	15	5	13	2	12	17	1	6	3	16	11		
a	1	4	8	5	6	9	11	14	15	20	7	13	12	19	3	18	17	10	2	16		
a	4	5	14	19	11	15	18	20	3	9	2	4	1	10	7	16	17	18	16	13	20	12
a	3	9	4	8	14	19	11	15	18	20	1	2	10	6	13	7	17	12	5	16		
a	11	15	6	9	14	19	2	8	4	5	18	20	3	7	1	17	10	12	13	16		
a	14	20	4	5	3	6	11	19	7	13	9	18	2	8	15	12	1	17	16	10		
a	10	17	8	5	13	7	15	18	11	14	1	9	12	16	4	19	20	2	3	6		
a	4	5	11	15	9	12	6	7	16	17	8	19	18	20	3	13	1	14				
a	3	7	14	19	11	15	4	20	1	17	10	13	6	18	9	12	11	15	5	16		
a	3	6	2	5	4	9	7	8	1	10	11	15	19	20	13	17	12	14	18	13		
a	3	8	1	19	2	18	4	5	6	18	7	13	8	9	10	7	11	1	14	16	20	15
a	4	5	8	6	9	3	12	6	17	1	11	15	16	18	13	19	20	14				
c	13	15	1	7	11	21	4	5	3	14	22	12	19	16	2	8	6	20	10	18	17	9
c	1	18	2	11	3	19	4	13	5	20	6	15	7	17	8	14	9	22	10	21		
c	3	19	21	22	4	5	14	16	6	8	1	2	11	12	13	15	9	17	18	20	7	10
c	5	14	2	21	8	14	5	16	6	10	1	13	4	7	8	18	12	19	15	17	3	9
c	5	16	9	21	3	22	2	11	12	7	4	14	8	20	17	7	6	15	10	19	1	15
c	1	12	2	14	3	19	4	7	5	16	13	18	6	9	10	15	9	21	11	22	8	17
c	14	17	13	15	8	18	21	22	2	12	4	16	3	20	6	10	11	7	5	9	1	19
c	6	15	1	5	7	17	14	8	16	10	4	18	3	20	2	12	18	22	21	9	11	13
c	3	9	4	5	1	7	21	22	11	6	10	16	13	15	2	12	8	18	19	20	14	17
c	7	14	19	3	18	15	21	22	17	1	10	6	2	20	11	12	16	9	3	8	4	5
c	14	18	13	15	1	4	19	20	12	22	10	16	3	9	2	21	6	7	5	8		
d	14	4	3	9	10	11	1	18	13	20	11	5	19	17	16	12	15	8	7	6		
d	4	5	7	20	14	17	13	10	12	16	15	19	8	18	3	9	2	6	1	11		
d	10	13	6	17	3	20	1	7	9	14	4	5	12	16	2	11	15	19	8	18		
d	5	9	2	7	3	10	18	10	15	19	20	6	17	9	1	20	5	13	11	2	6	20
d	1	20	2	11	3	17	4	13	5	14	6	11	7	14	8	15	9	14	10	17	11	2
d	11	12	15	8	13	4	14	9	6	20	7	1	2	19	3	10	18	16	5	17	19	11
d	2	7	9	14	19	11	4	5	14	17	9	10	12	16								
d	16	8	1	20	9	17	10	13	12	19	2	4	5	7	3	14	6	15	18	11		
d	14	17	12	16	5	13	8	15	9	11	1	10	2	7	4	19	3	20	6	18		
d	6	20	9	14	3	17	4	5	8	16	15	19	11	2								
d	1	7	4	5	9	10	11	15	3	6	2	13	8	18	14	17	16	19	12	20		
d	3	20	5	13	1	14	8	18	6	17	9	12	4	19	7	16	2	15	10	11		
d	19	15	2	20	11	7	1	14	5	12	3	10	4	13	6	17	8	16	9	18		
d	7	11	3	17	10	13	9	20	5	14	8	15	2	19	1	18	4	6	12	16		

Array no																						
b	5	6	1	7	4	9	2	10	3	8												
b	10	3	14	15	20	13	1	6	7	5	9	18	4	16								
b	2	12	3	14	6	15	10	13	11	20	1	19	4	16	4	7	17	18	8	9		
b	1	2	3	6	4	12	5	8	7	8	9	12	10	13	11	20	16	18	14	19	13	20
b	10	13	7	5	1	6	3	14	2	12	15	19	9	18	4	17	8	16	11	20		
b	3	14	13	10	8	19	20	6	11	18	2	12	7	5	16	17	13	19	8	14	9	4
b	12	19	2	10	5	6	15	8	1	7	11	9	4	17	13	16	14	20	3	18		
b	6	13	13	12	10	19	14	3	20	9	2	8	4	7	11	17	18	1	4	15		
b	12	14	10	11	8	16	7	19	6	18	1	2	13	3	17	9	15	5	4	20		
b	4	9	19	11	8	14	5	21	1	15	10	18	9	12	3	22	16	17	6	2	13	20
b	1	6	7	12	13	14	15	19	18	11	16	17	2	3	5	8	4	10	9	20		
b	3	10	11	20	2	12	1	6	7	8	19	14	17	5	9	16	15	13	18	4		
b	3	6	9	18	8	14	5	20	19	21	12	13	4	10	17	15	2	17	1	16	11	22
b	1	6	1	2	3	5	4	10	2	8	3	9	7	9								
b	1	2	4	9	18	22	19	21	3	6	20	16	15	8	7	12	14	17	11	10	5	13
b	16	17	5	6	3	19	2	8	1	15	9	18	4	13	11	22	12	20	14	21	7	10
b	4	8	1	12	10	22	8	3	17	21	15	16	2	20	5	6	7	9	19	3	3	12
b	3	6	1	15	2	5	11	22	8	12	10	18	4	9	14	21	7	9	21	6	20	16
b	6	19	3	21	7	10	1	8	12	17	2	5	4	9	13	15	14	16	18	20	11	22
b	8	20	1	10	9	17	19	11	7	18	5	15	13	2	6	3	4	16	12	22	14	21
b	18	22	4	10	6	19	3	4	1	13	16	17	2	20	5	11	7	9	8	12	21	15
b	8	17	13	22	19	21	2	8	1	15	7	10	11	18	6	19	5	16	12	16	4	9
b	6	5	9	10	16	17	11	12	3	14	19	18	1	13	20	4	15	2	7	8		
b	1	7	2	6	3	4	5	10	9	2												
b	1	3	2	15	4	11	5	18	6	4	7	16	8	19	9	13	10	17	11	20		
b	16	15	8	13	1	10	11	17	12	18	19	6	7	20	14	4	2	5	3	7		
b	1	11	10	13	2	12	4	5	14	3	17	20	6	19	15	7	16	8	18			
g	2	6	1	7	5	17	9	12	19	14	15	16	13	4	3	11	5	15	10	20		
g	5	16	9	12	10	11	6	19	3	14	1	13	8	20	2	18	7	15	14	17		
g	2	19	5	16	12	18	4	20	1	8	10	11	6	9	17	13	15	17	3	14		
g	19	2	18	12	9	15	5	6	17	11	7	1	10	14	20	16	13	4	8	3		
g	1	7	2	11	3	19	4	20	5	16	12	15	13	8	10	18	6	9	14	17		
g	3	5	7	18	15	18	4	9	6	6	13	2	12	17	1	8	10	14	11	20		
h	6	7	3	4	16	18	14	15	19	20	11	12	2	8	9	10	17	18	1	2	5	9
h	2	6	3	10	7	4	14	18	11	19												
h	1	9	2	17	3	15	4	16	5	19	6	14	7	11	8	20	9	3	10	20	11	2
h	5	10	6	7	3	4	13	14	15	20	8	11	2	3	9	17	12	20	19	11	18	12
h	4	16	4	14	3	17	7	16	13	18	11	19	12	15								
h	7	16	4	18	14	5	8	11	6	15	9	19	10	2	12	17	3	13	1	20		
h	16	13	1	15	3	18	10	11	4	16	2	11	5	14	8	15	9	19	10	12	10	17
h	7	16	6	14	4	9	1	8	2	18	5	10	3	6	12	20	11	19	13	15		
h	7	16	6	19	4	18	3	9	1	8	5	10	11	17	2	13	14	15	12	20		
h	4	14	3	13	9	11	5	19	2	15	6	17	1	12	8	20	7	16	10	18		
f	1	20	4	11	6	2	8	17	16	12	15	14	3	9	18	19	7	5	13	10		
f	2	12	6	14	1	20	8	13	7	11	5	18	17	19	9	16	3	15	10	1	4	17
f	1	8	2	6	7	13	3	20	12	19	9	15	4	5	10	18	11	17	16	14		
f	2	20	3	9	4	12	5	10	6	14	7	20	8	16	9	3	10	5	11	17		
f	2	6	17	8	3	9	7	11	1	2	13	16	12	14	4	15	18	19	5	10		
f	3	9	4	5	15	18	6	2	13	15	11	17	7	8	2	12	14	18	1	10		
f	9	15	6	19	2	13	6	18	8	12	10	17	3	16	5	20	4	11	7	14		
f	7	8	7	18	3	9	16	7	6	20	2	10	4	5	13	20	6	19	11	5	13	8
f	2	6	18	19	7	8	13	20	12	14	15	16	4	5	3	9	11	17	1	10		
f	2	6	3	9	1	12	13	20	8	17	14	15	16	19	4	5	10	18	11	7		
f	1	12	16	13	18	15	2	14	4	5	8	19	6	7	17	20	3	9	10	11		
f	2	14	8	17	1	20	7	16	10	15	5	18	6	11	9	13	4	19	12	3		
f	2	6	13	20	1	8	3	9	11	12	16	19	5	18	4	7	10	15	14	17		
f	1	17	6	12	7	11	2	20	9	15	13	18	4	10	8	14	7	13	2	7	13	20

Study 3: Similarity rating task

The first 9 columns (in each of the three sets across the page) represent faces in the array, in the order in which they appear (excluding the target).

5	7	3	2	8	7	2	6	6	high	discont	1	1	1	8	1	1	7	1	2	1	low	discont	2	2	3	1	1	3	2	1	2	3	high	contin	2
0	4	2	2	3	6	0	0	0	high	discont	1	3	1	4	5	7	2	4	1	5	mod	contin	2	6	2	1	1	2	2	3	1	2	low	discont	2
4	0	0	6	0	0	0	0	0	mod	discont	1	2	5	4	8	9	7	4	5	3	high	contin	1	3	1	2	3	2	4	6	1	1	mod	contin	1
6	0	0	2	0	1	0	1	5	high	contin	2	3	4	7	3	9	8	7	5	4	low	discont	2	6	7	1	3	0	8	1	2	4	high	discont	2
0	4	5	0	5	1	0	0	5	low	discont	2	4	9	6	4	0	5	3	2	5	mod	contin	2	7	2	7	8	3	5	3	4	4	low	contin	2
5	0	0	0	1	0	1	0	0	mod	contin	1	3	4	4	5	4	4	5	3	4	high	contin	1	4	2	0	3	2	8	1	7	1	mod	discont	2
3	0	0	.	0	5	0	4	4	high	contin	2	3	4	5	4	4	4	5	3	2	low	discont	2	2	3	1	0	3	2	0	1	3	high	contin	2
0	0	7	0	4	0	2	1	4	low	discont	2	4	3	3	3	6	5	4	2	6	mod	contin	2	3	2	4	1	5	2	2	1	2	low	discont	2
3	0	2	2	1	1	4	.	0	mod	contin	1	7	4	0	3	0	6	3	3	9	high	discont	2	2	1	0	1	1	2	6	1	0	mod	contin	1
7	3	0	3	0	4	0	0	3	high	discont	2	4	5	7	7	3	2	6	7	3	low	contin	2	7	0	3	1	1	7	0	5	4	high	contin	2
2	0	1	1	1	0	0	0	0	low	contin	2	7	3	0	4	4	6	0	4	4	mod	discont	2	0	1	5	0	5	2	4	4	5	low	discont	2
1	1	0	0	0	6	0	1	0	mod	discont	2	2	5	4	5	4	3	2	4	1	high	contin	1	5	0	4	7	0	0	3	0	0	mod	contin	1
2	4	2	5	2	5	5	2	6	high	contin	1	1	3	3	1	2	6	1	1	0	low	discont	2	6	7	0	5	0	5	0	1	7	high	discont	2
2	4	6	2	2	7	6	2	2	low	discont	2	2	0	5	1	4	3	5	2	7	mod	contin	2	3	0	0	2	5	0	0	4	1	low	contin	2
7	2	2	4	4	7	6	2	7	mod	contin	2	3	4	3	3	1	2	3	3	3	high	contin	1	0	4	0	0	6	7	0	0	0	mod	discont	2
2	0	.	6	0	8	0	5	8	high	contin	2	3	3	2	2	4	5	4	2	2	low	discont	2	1	3	2	1	1	4	0	0	0	high	contin	1
2	0	3	0	5	0	8	5	10	low	discont	2	2	0	2	2	2	1	4	3	3	mod	contin	2	1	0	1	0	2	0	3	2	3	mod	contin	2
.	5	7	8	0	0	6	2	0	mod	contin	1	4	6	0	7	1	3	2	2	3	high	discont	2	1	2	2	4	3	3	2	4	2	high	contin	1
4	7	3	6	2	5	6	4	7	high	contin	1	4	0	0	3	4	.	4	5	2	low	contin	2	1	1	4	2	2	5	3	1	0	low	discont	2
4	7	6	4	6	7	6	4	3	low	discont	2	4	2	4	2	3	2	2	0	0	mod	discont	2	2	1	3	1	4	3	2	0	5	mod	contin	2
5	2	4	5	6	5	4	3	8	mod	contin	2	4	1	0	2	0	4	2	0	0	high	contin	2	3	10	2	6	4	8	5	7	9	high	discont	1
0	5	1	0	3	7	5	8	4	high	discont	1	0	2	0	0	0	0	0	0	4	low	discont	2	6	5	10	4	7	8	9	3	2	low	contin	1
0	2	3	5	2	3	0	1	4	low	contin	1	2	0	0	1	2	2	2	0	0	mod	contin	1	8	7	4	10	6	5	2	3	9	mod	discont	1
1	1	0	1	0	0	0	1	1	mod	discont	1	4	2	1	0	2	6	2	3	7	high	contin	2	3	5	2	1	7	8	4	6	7	high	discont	1
2	2	0	4	0	1	0	0	0	high	discont	2	3	3	2	0	7	5	8	3	9	low	discont	2	4	3	7	5	4	5	6	3	4	low	contin	1
4	0	3	2	1	4	0	0	2	low	contin	2	3	3	3	2	3	0	5	1	1	mod	contin	1	8	5	4	3	5	1	2	6	2	mod	discont	1
1	4	0	1	3	4	1	2	2	mod	discont	2	7	2	1	0	2	6	0	4	4	high	contin	2	2	4	2	3	4	7	4	4	7	high	discont	1
4	6	1	4	1	6	3	2	5	high	discont	2	1	2	0	.	9	5	5	2	3	low	discont	2	4	3	7	5	4	3	3	3	4	low	contin	1
2	1	6	7	1	5	3	2	3	low	contin	2	6	3	1	0	5	6	7	0	0	mod	contin	1	7	2	2	3	4	2	2	4	2	mod	discont	1
2	3	2	2	2	7	2	4	4	mod	discont	2	6	6	1	4	10	6	9	10	7	high	discont	2	0	5	2	1	7	7	1	2	2	high	discont	1
7	4	3	1	5	4	2	5	2	high	discont	2	7	4	8	6	7	9	9	9	8	low	contin	2	2	1	6	4	4	2	6	1	1	low	contin	1
7	4	7	6	4	7	4	4	5	low	contin	2	6	7	10	10	8	9	5	8	8	mod	discont	2	5	6	1	1	1	0	0	2	0	mod	discont	1
8	6	2	6	6	4	4	4	5	mod	discont	2	0	4	2	4	4	5	7	0	0	low	contin	1	5	3	5	4	6	8	4	6	6	high	discont	1
7	7	1	5	0	8	2	1	5	high	discont	2	0	3	4	10	10	0	0	0	0	mod	discont	1	1	0	4	6	1	0	1	0	0	low	contin	1
3	3	7	6	5	3	3	4	7	low	contin	2	7	8	1	6	0	9	4	3	7	high	discont	2	5	7	1	6	1	0	0	2	4	mod	discont	1
3	4	0	3	5	6	2	5	3	mod	discont	2	4	1	4	7	6	1	2	6	5	low	contin	2	3	3	2	2	0	3	2	2	2	high	contin	1
5	5	5	5	2	4	4	4	5	high	discont	1	2	7	0	1	4	9	0	1	0	mod	discont	2	2	4	8	2	3	6	5	3	4	low	discont	2
0	0	3	7	1	2	0	0	2	low	contin	1	0	2	0	0	0	0	0	0	0	high	discont	1	7	2	0	1	4	6	6	6	8	mod	contin	2
2	2	0	3	2	0	4	3	4	mod	discont	1	4	4	4	2	4	4	3	3	3	low	contin	1	5	4	1	2	1	6	6	4	7	high	discont	2
3	7	4	6	5	6	4	3	1	high	contin	1	3	5	3	3	3	0	0	0	0	mod	discont	1	8	7	9	8	7	6	8	8	3	low	contin	2
0	1	2	1	3	6	3	2	1	low	discont	2	2	3	0	2	1	4	2	2	5	high	discont	2	9	8	2	5	6	7	1	1	1	mod	discont	2
3	0	6	1	8	4	1	0	8	mod	contin	2	4	0	2	2	0	5	2	3	0	low	contin	2	6	5	1	4	5	9	4	6	8	high	discont	2
0	0	0	3	4	3	0	4	0	high	contin	1	2	1	1	2	2	1	1	1	1	mod	discont	2	5	3	6	7	4	8	1	7	9	low	contin	2
0	1	2	0	3	6	2	0	0	low	discont	2	1	2	1	.	3	5	2	2	3	high	discont	1	7	6	3	4	5	3	2	7	1	mod	discont	2
3	0	5	1	7	1	6	0	9	mod	contin	2	2	1	4	3	2	3	4	0	0	low	contin	1	1	4	0	6	1	3	2	1	5	high	contin	1
6	10	8	7	7	8	3	2	1	high	contin	1	2	1	1	1	1	0	0	2	1	mod	discont	1	2	1	4	2	0	8	7	4	1	low	discont	2
5	8	9	4	8	9	2	2	1	low	discont	2	6	4	2	5	0	4	3	1	8	high	discont	2	1	0	3	0	5	0	6	2	7	mod	contin	2
7	1	2	7	5	5	6	4	5	mod	contin	2	7	3	4	6	3	6	4	7	3	low	contin	2	2	6	2	5	3	6	4	0	4	high	contin	1
8	7	2	3	4	5	2	2	6	high	discont	2	2	5	0	2	4	7	0	7	4	mod	discont	2	3	3	7	2	2	.	1	6	4	low	discont	2
5	3	4	7	5	2	2	3	2	low	contin	2	6	2	2	2	5	6	2	5	2	high	contin	2	2	1	7	4	3	2	1	2	8	mod	contin	2
2	3	1	1	3	8	1	4	2	mod	discont	2	2	6	7	2	6	5	1	1	3	low	discont	2	2	1	0	4	0	0	1	1	1	high	contin	1
1	4	4	4	0	0	0	1	2	high	contin	1	6	4	2	6	6	2	7	2	4	mod	contin	1	0	0	1	1	0	4	2	3	0	low	discont	2
2	0	1	1	2	5	5	0	2	low	discont	2	0	1	1	0	0	4	0	2	6	high	contin	2	5	2	1	0	2	1	2	0	1	mod	contin	2
2	1	0	1	0	2	4	1	3	mod	contin	2	0	3	1	0	2	0	6	5	4	low	discont	2	0	2	4	3	7	6	0	0	3	high	discont	1
4	6	0	6	0	8	.	3	7	high	discont	2	3	0	1	1	1	0	4	0	0	mod	contin	1	0	3	2	5	4	2	1	0	0	low	contin	1
6	0	3	1	5	1	2	7	1	low	contin	2	4	2	1	0	1	3	2	4	4	high	contin	2	3	0	0	6	3	0	8	5	.	mod	discont	1
1	1	0	1	1	3	0	1	0	mod	discont	2	3	2	1	2	5	4	2	3	3	low	discont	2	0	6	0	3	5	7	3	1	4	high	contin	1
3	2	0	3	0	8	1	0	7	high	discont	2	0	1	1	2	2	1	5	0	1	mod	contin	1	0	2	1	1	1	6	3	1	0	low	discont	2
1	0	0	0	0	0	1	0	0	low	contin	2	1	3	1	1	7	8	4	5	3	high	discont	1	2	0	3	2	7	6	1	0	4	mod	contin	2
0	2	0	0	0	3	0	0	0	mod	discont	2	2	1	4	0	1	0	0	1	0	low	contin	1	3	4	1	2	7	9	8	5	6	high	discont	1
6	7	2	9	4	10	3	5	8	high	discont	2	7	1	2	4	2	0	0	1	0	mod	discont	1	1	4	6	5	7	8	9	2	3	low	contin	1
8	7	2	4	10	6	5	9	3	low	contin	2	1	4	3	2	1	3	2	0	0	high	contin	1	9	6	3	5	8	2	1	7	4	mod	discont	1
9	4	7	3	5	10	2	8	6	mod	discont	2	0	1	2	3	0	5	2	0	0	low	discont	2	8	1	1	0	0	3	2	3	6	high	contin	2
1	2	0	3	3	2	1	0	0	high	discont	2	2	0	1	1	1	0	6	2	7	mod	contin	2	3	3	5	5	6	2	6	4	8	low	discont	2
1	0	5	4	2	0	1	6	3	low	contin	2	4	3	0	2	2	4	3	3	2	high	contin	2	9	2	8	4	3	5	9	2	2	mod	contin	1
0	1	0	0	1	3	0	1	0	mod	discont	2	0	1	2	0	3	0	0	0	2	low	discont	2	7	.	0	4	2	3	0	3	5	high	discont	2
1	6	3	5	3	6	2	3	2	high	contin	1	5	0	0	0	2	5	3	0	0	mod	contin	1	0	6	0	2	1	3	0	1	1	mod	discont	2
1	2	7	3	6	6	2	4	3	low	discont	2	5	2	0	0	2	8	2	2	4	high	contin	2	9	1	0	0	1	7	0	0	2	high	contin	2
1	0	7	1	6	6	1	0	7	mod	contin	2	0	0	5	2	2	0	3	4	7	low	discont	2	0	0	6	0	0	0	8	1	7	low	discont	2
1	6	1	6	1	1	0	1	3	high	contin	1	5	6	4	7	5	5	5	0	0	mod	contin	1	8	0	6	7	0	2	8	0	0	mod	contin	1
0	2	0	0	0	4	1	1	0	low	discont	2	3	2	1	2	1	5	1	2	4	high														

Study 3: Distinctiveness rating task

Notation of the form f1, f2 .. f28, heading columns, represent face images in the rated array. The subscripts indicate the order in which they appear in the array.

Array no	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28
6	7	1	1	1	15	7	1	1	7	7	15	7	1	1	15	7	15	7	1	1	1	1	15	1	7	1	1	1
6	10	10	6	12	6	10	8	6	7	7	12	7	11	12	11	6	12	7	10	6	7	6	12	6	4	5	7	5
6	2	13	2	7	3	1	2	2	4	1	14	10	2	2	2	3	8	3	2	2	1	4	12	4	12	3	10	5
6	7	7	7	1	1	1	15	15	1	1	1	1	7	1	1	7	7	7	7	7	1	1	7	7	1	1		1
6	7	15	7	7	7	1	15	7	1	1	1	1	7	1	1	1	1	1	1	7	1	1	7	7	1	1	7	1
6	7	15	7	1	7	1	7	1	7	1	7	7	7	1	15	1	7	1	7	7	1	15	15	1	7	7	1	1
6	7	15	1	1	7	15	7	7	7	1	15	7	7	1	7	1	7	1	1	1	1	7	7	1	15	7	7	7
6	4	4	7	5	12	10	7	5	7	6	10	4	3	3	3	3	3	3	2	2	2	5	6	7	8	4	3	4
6	6	4	4	10	15	12	4	12	2	4	15	7	3	2	10	9	13	2	2	4	2	5	10	2	10	1	6	10
6	1	7	1	1	7	7	1	1	7	1	15	7	1	7	15	1	15	7	7	1	1	1	7	7	7	1	1	1
7	5	9	7	5	15	12	1	10	15	8	8	9	4	12	9	10	11	7	15	8	8	9	15	9	10	14	12	8
7	7	1	1	7	7	1	7	7	1	1	7	1	7	7	1	7	1	1	1	7	1	1	1	7	15	1	1	15
7	6	7	6	15	6	6	8	2	7	4	4	3	15	5	5	7	5	8	3	3	8	5	3	4	9	7	5	7
7	1	1	7	7	7	1	1	7	7	1	7	1	7	7	1	1	1	7	7	1	1	1	1	7	7	1	7	7
7	2	5	5	7	12	8	6	10	14	10	11	12	10	12	10	6	7	7	10	7	5	8	8	4	6	8	12	10
7	1	1	1	15	7	1	7	15	15	15	7	7	15	7	1	7	7	7	7	7	7	7	1	7	15	1	7	7
7	1	7	1	1	7	1	1	7	1	1	7	1	15	1	1	15	1	1	1	1	1	1	1	1	15	1	1	7
10	8	1	9	1	5	1	1	7	5	3	2	2	1	5	2	2	15	1	6	7	15	2	1	2	5	7		
10	5	5	2	4	8	8	8	10	9	5	4	3	3	2	6	8	9	4	3	5	8	9	8	9	9	7		
10	3	3	5	4	6	3	3	6	4	2	4	4	2	6	7	7	7	2	2	7	8	2	2	2	7	5		
10	7	7	1	1	15	5	7	1	7	1	7	7	15	15	7		15	7	15	7	1	7	7	15	15	1		
10	3	7	7	3	9	8	8	5	7	5	4	4	4	4	4	3	8	3	7	7	9	4	4	4	4	5		
10	5	7	15	7	7	6	7	6	7	6	6	10	6	10	4	4	15	12	3	3	13	15	12	6	13	6		
10	2	5	7	1	3	3	9	1	1	2	7	2	2	5	5	5	4	6	7	4	12	5	6	7	5	4		
10	13	7	5	11	7	6	9	6	7	7	7	15	7	11	10	15	15	13	6	13	15	7	7	7	10	9		
10	8	3	5	10	5			8	8	5	3	10	7	5	12	12	15	9	7	5	15	4	2	4	7	6		

Study 4: Similarity rating task, across frontal and profile views

Notation of the form f1, f2 .. f28, heading columns, represent face images in the rated array. The subscripts indicate the order in which they appear in the array. Hi = high distinctiveness; lo = low distinctiveness; fr = frontal view; pr = profile view; fp = frontal + profile view.

F1	F2	F3	F4	F5	F6	F7	F8	F9	Distinct	View	Seq	F1	F2	F3	F4	F5	F6	F7	F8	F9	Distinct	View	Seq	F1	F2	F3	F4	F5	F6	F7	F8	F9	Distinct	View	Seq
5	0	2	3	2	0	1	2	2	hi	fp	2	6	9	2	3	0	0	4	7	10	lo	fr	2	2	2	2	0	7	9	7	4	1	lo	fp	1
5	6	4	5	5	3	3	3	5	hi	fp	1	0	0	0	4	4	5	6	7	0	hi	fr	1	0	0	2	0	0	0	1	3	0	hi	fp	2
5	5	5	6	5	4	4	4	8	lo	fp	2	3	0	0	0	3	5	1	4	0	lo	pr	1	0	0	2	0	2	0	0	0	0	lo	fp	1
3	2	5	2	3	5	2	1	3	hi	pr	1	3	3	3	8	5	5	4	8	3	hi	fp	2	0	0	5	0	3	0	3	0	6	hi	fr	2
2	3	2	5	4	2	2	2	7	lo	fr	1	2	1	4	1	7	7	5	2	4	lo	fp	1	0	3	2	0	1	2	0	4	0	lo	pr	2
6		7	0	3	4	0	1	1	hi	pr	1	0	0	0	3	2	2	3	0	0	hi	fr	2	0	1	0	1	2	0	2	0	2	hi	fr	2
0	5	0	4	2	1	0		5	lo	fr	1	0	2	0	0	2	2	0	4	2	lo	pr	2	0	1	2	0	1	0	0	2	0	lo	pr	2
5	6	9	1	3	8	1	0	1	hi	pr	1	0	0	0	1	0	0	1	0	0	hi	fp	2	3	1	2	0	4	4	3	0	1	hi	pr	1
0	4	1	1	1	3	2	1	1	lo	fr	1	0	0	2	0	3	0	0	0	0	lo	fp	1	0	4	0	4	2	0	1	1	3	lo	fr	1
0	5	10	0	9	7	8	0	5	hi	pr	1	3	5	4	2	6	2	2	7	5	hi	fr	2	3	5	5	6	1	2	6	0	5	hi	fr	2
5	8	0	1	1	0	8	7	9	lo	fr	1	7	8	6	4	6	6	7	9	7	lo	pr	2	1	5	6	3	4	8	3	7	2	lo	pr	2
1	3	3	0	4	2	6	2	2	hi	pr	1	3	5	1	1	6	8	7	6	4	hi	pr	2	5	0	10	0	10	5	0	0	5	hi	pr	2
0	3	1	1	1	1	1	1	6	lo	fr	1	2	8	0	2	4	3	0	6	7	lo	fr	2	0	0	0	0	0	0	10	10	0	lo	fr	2
1	0	5	0	4	2	2	2	5	hi	pr	1	2	0	0	0	0	2	0	0	0	hi	fr	1	1	3	0	1	3	1	6	6	0	hi	fp	2
1	1	0	0	1	0	1	0	0	lo	fr	1	3	2	1	0	1	1	0	3	0	lo	pr	1	2	0	4	0	8	2	7	0	2	lo	fp	1
5	5	4	3	5	8	4	3	7	hi	pr	1	2	3	6	2	7	5	7	9	8	hi	pr	2	1	1	2	3	1	1	4	2	2	hi	fr	2
2	7	2	5	6	3	6	4	5	lo	fr	1	3	5	1	2	1	2	4	8	9	lo	fr	2	0	2	4	2	2	3	2	5	2	lo	pr	2
1	1	1	0	1	2	5	1	2	hi	pr	1	0	5	2	3	0	1	1	0	4	hi	fp	1	4	0	0	1	0	1	6	0	6	hi	fr	2
0	3	0	2	2	0	3	0	0	lo	fr	1	0	2	3	4	5	0	1	0	5	lo	fp	2	0	6	2	0	3	3	4	8	0	lo	pr	2
4	0	6	0	6	5	0	1	6	hi	pr	1	0	0	0	5	1	3	6	0	0	hi	fr	1	3	4	3	6	2	4	4	3	4	hi	fr	2
0	0	1	6	0	2	2	1	3	lo	fr	1	6	1	0	0	0	4	2	7	0	lo	pr	1	2	2	4	2	1	3	3	5	3	lo	pr	2
10	5	4	8	5	3	3	3	5	hi	fr	1	2	4	7	4	2	8	5	2	5	hi	fr	2	3	2	5	3	3	3	5	2	3	hi	fr	2
7	8	3	2	3	2	2	5	2	lo	pr	1	5	2	6	0	4	8	7	7	6	lo	pr	2	1	1	6	5	5	4	5	10	5	lo	pr	2
3	0	0	4	5	0	1	0	3	hi	fr	1	1	0	10	0	4	9	8	1	7	hi	pr	1	3	2	7	6	4	7	7	6	5	hi	fr	2
1	4	0	1	0	1	2	2	0	lo	pr	1	8	1	0	1	9	3	1	2	0	lo	fr	1	4	5	9	8	8	7	7	8	6	lo	pr	2
9	6	4	7	7	6	2	3	4	hi	fr	1	7	4	6	4	6	7	7	6	3	hi	fp	1	3	5	8	0	6	9	2	4	1	hi	pr	2
6	6	4	4	5	6	7	8	9	lo	pr	1	7	8	4	7	7	9	9	8	7	lo	fp	2	3	6	0	1	3	0	7	2	4	lo	fr	2
4	6	2	5	6	3	4	3	4	hi	fr	1	0	0	1	0	0	1	0	0	3	hi	pr	2	4	3	7	2	3	4	6	5	6	hi	pr	2
4	5	2	3	3	4	3	6	2	lo	pr	1	5	6	1	1	0	0	4	1	0	lo	fr	2	5	7	3	4	3	2	4	7	6	lo	fr	2
1	0	4	0	2	0	1	2	5	hi	fr	2	2	3	7	4	4	8	2	3	6	hi	pr	2	0	0	2	0	0	0	0	0	8	hi	pr	2
4	5	3	2	2	7	0	8	1	lo	pr	2	3	5	2	2	1	3	7	8	7	lo	fr	2	0	0	0	0	0	0	4	0	0	lo	fr	2
2	3	3	0	3	2	3	0	3	hi	pr	1	0	0	0	0	0	7	5	0	0	hi	fr	2	10	4	3	2	8	7	5	6	9	hi	pr	2
0	1	0	1	2	1	1	0	0	lo	fr	1	0	0	8	0	0	0	0	5	0	lo	pr	2	5	8	2	4	3	5	10	7	6	lo	fr	2
2	7	9	8	10	3	5	6	4	hi	fp	2	7	8	8	9	7	5	10	7	6	hi	fr	2	0	3	6	6	5	1	8	0	4	hi	fr	2
0	0	9	0	10	7	8		6	lo	fp	1	1	7	7	7	8	4	4	8	9	lo	pr	2	1	6	8	1	4	7	2	7	0	lo	pr	2
1	5	3	6	5	3	6	1	7	hi	fp	1	7	4	6	5	3	8	9	4	10	hi	fr	2	10	10	9	2	4	0	8	10	10	hi	fp	1
6	6	2	5	2	0	0	2	4	lo	fp	2	2	4	9	7	8	5	3	10	6	lo	pr	2	5	10	10	4	8	10	7	10	8	lo	fp	2
0	0	1	0	3	2	0	0	0	hi	pr	2	5	5	0	2	6	3	4	0	0	hi	pr	2	1	8	6	1	1	3	7	5	5	hi	pr	2
4	1	0	0	0	0	0	1	0	lo	fr	2	0	6	0	0	2	0	3	6	6	lo	fr	2	6	7	0	5	1	1	4	7	5	lo	fr	2
6	4	2	7	7	4	4	5	4	hi	fr	1	0	2	8	8	1	4	7	0	2	hi	fr	2	2	5	3	1	6	5	2	2	3	hi	pr	2
8	3	2	3	4	6	6	6	2	lo	pr	1	1	5	6	1	3	5	2	4	2	lo	pr	2	2	2	1	1	1	2	6	5	2	lo	fr	2
2	2	6	4	5	2	1	2	1	hi	fp	2	1	3	1	4	3	0	0	0	0	hi	fp	2	0	6	4	3	6	5	6	4	3	hi	pr	2
1	2	3	2	4	2	2	1	1	lo	fp	1	2	0	3	0	5	3	5	0	2	lo	fp	1	1	4	5	4	5	6	8	7	4	lo	fr	2
5	4	4	4	1	1	3	3	5	hi	fr	1	0	1	4	4	1	0	0	0	1	hi	fr	2	5	3	4	2	6	7	6	3	5	hi	pr	1
4	5	1	1	2	4	5	5	3	lo	pr	1	0	3	5	1	1	1	0	6	1	lo	pr	2	3	3	2	5	5	2	3	3	4	lo	fr	1
4	5	2	6	6	5	2	2	5	hi	fp	2	2	1	2	0	1	3	1	0	3	hi	pr	2	0	0	3	2	1	1	0	0	0	hi	fp	2
4	2	5	3	6	2	3	5	4	lo	fp	1	2	2	0	2	1	0	1	4	4	lo	fr	2	7	2	3	2	2	1		0	0	lo	fp	1
1	8	1	5	5	2	4	4	1	hi	fp	2	3	8	2	7	5	2	5	6	3	hi	fp	2	9	7	7	9	9	8	6	6	7	hi	fp	2
3	3	5	3	8	4		2	6	lo	fp	1	5	2	7	4	8	7	6	4	7	lo	fp	1	8	6	6	6	5	8	6	5		lo	fp	1
8	5	4	5	4	4	6	7	4	hi	fr	1	3	1	3	4	5	2	1	1	3	hi	fr	1	1	3	1	0	6	2	2	1	2	hi	pr	2
5	6	2	6	3	6	4	8	3	lo	pr	1	2	5	1	2	3	3	4	5	1	lo	pr	1	8	4	1	2	0	3	2	5	4	lo	fr	2
7	2	1	5	6	5	2	3	3	hi	fr	1	3	1	0	2	5	3	1	1	0	hi	fp	2	6	4	7	4	7	5	9	6	2	hi	fr	2
7	6	5	5	5	6	5	5	4	lo	pr	1	1	0	1	1	1	2	3	1	5	lo	fp	1	4	3	5	2	3	4	5	4	2	lo	pr	2
7	7	5	2	1	7	8	0	4	hi	fr	2	4	4	8	4	5	4	2	1	4	hi	pr	1	2	4	4	7	3	1	7	0	8	hi	fr	2
4	1	8	6	8	2	3	8	5	lo	pr	2	1	2	2	3	3	1	2	2	4	lo	fr	1	3	5	6	4	5	4	3	9	3	lo	pr	2
5	1	6	2	7	4	6	2	7	hi	pr	1	3	4	3	6	2	2	5	3	4	hi	fp	2	4	4	2	2	6	7	5	4	3	hi	pr	1
2	6	2	5	1	2	5	3	5	lo	fr	1	2	1	3	1	5	2	3	1	7	lo	fp	1												
5	2	4	3	3	5	2	2	1	hi	fp	2	5	3	0	6	2	1	0	0	0	hi	fp	2												
1	1	3	2	2	2	4	3	3	lo	fp	1	2	0	6	0	6	5	5	0	6	lo	fp	1												
0	0	2	0	9	9	3	3	10	hi	pr	2	2	2	2	7	5	3	3	3	5	hi	fp	2												

Study 5: Test-retest reliability of subject ratings of similarity

Initial = array presented at initial stage of experiment; Follow = array presented at final stage of experiment. 't1f1' = first face in array presented at initial stage, 't2f1' = first face in array presented at final stage, and so on.

Subject	Initial	t1f1	t1f2	t1f3	t1f4	t1f5	t1f6	t1f7	t1f8	t1f9	Follow	t2f1	t2f2	t2f3	t2f4	t2f5	t2f6	t2f7	t2f8	t2f9
1	t301	10	30	10	50	10	10	20	15	30										
2	t301	10	50	20	60	40	30	50	30	0	t302	60	40	40	0	70	0	70	60	60
3	t301	30	40	20	50	50	60	20	70	60	t302	90	50	40	0	60	5	35	50	80
4	t301	30	50	20	40	40	60	70	30	50										
5	t301	40	70	30	50	30	30	50	30	70	t302	70	30	30	50	50	30	80		70
6	t301	0	20	0	30	0	20	30	10	20	t303	10	5	0	15	0	70	10		5
7	t301	0	0	30	40	60	50	0	50	0										
8	t301	45	65	25	80	30	35	50	30	55	t302	90	55	50	30	75	60	80	65	70
9	t301	0	30	10	0	0	0	0	0	0	t302	60	0	20	0	40	0	30	10	0
10	t301	0	50	0	50	0	0	0	0	0	t302	50	0	0	0	0	0	0	0	0
11	t301	20	60	10	75	30	40	60	10	20	t303	80	20	20	10	50	20	60	60	50
12	t301	20	60	0	45	10	10	0	5	5										
13	t301	40	80	50	80	45	70	45	40	50										
14	t301	0	80	0	30	40	20	10	20	0	t303	25	0	15	5	60	15	50	10	30
15	t301	20	40	30	40	30	40	45	30	40	t303	60	50	45	45	65	50	50		55
16	t301	60	90	70	70	80	70	60	80	60	t303	75	50	60	70	70	55	60	75	60
17	t301	0	70	20	30	10	75	25	25	30										
18	t301	10	75	10	5	5	50	75	10	50										
19	t301	5	10	0	0	0	5	0	0	0	t303	0	0	0	5	5	5	0	0	5

Study 6: Mock witness task

Dist = distinctiveness; Sim. = similarity; Mem. = array member chosen; S = subject; Crt. = identification correctness (1 = correct).

Dist	Sim	Mem	S	Crt	Dist	Sim	Mem	S	Crt	Dist	Sim	Mem	S	Crt	Dist	Sim	Mem	S	Crt	Dist	Sim	Mem	S	Crt
3	5	4	1	1	3	6	8	35	1	3	2	3	69	1	3	6	8	102	1	3	4	2	136	1
2	5	4	1	1	2	2	4	35	1	2	6	3	69	1	2	2	4	102	1	2	1	6	136	0
1	2	3	1	1	1	3	5	35	0	1	1	1	69	1	1	3	6	102	0	1	4	6	136	1
3	5	4	2	1	3	1	8	36	1	3	2	3	70	1	3	5	4	103	1	3	4	2	137	1
2	5	4	2	1	2	3	5	36	0	2	6	3	70	1	2	5	4	103	1	2	1	1	137	1
1	2	3	2	1	1	6	2	36	1	1	1	5	70	0	1	2	2	103	0	1	4	6	137	1
3	2	3	3	1	3	6	8	37	1	3	6	8	71	1	3	1	3	104	0	3	3	4	138	1
2	6	3	3	1	2	2	4	37	1	2	2	4	71	1	2	3	2	104	1	2	4	8	138	1
1	1	3	3	0	1	3	1	37	0	1	3	4	71	1	1	6	2	104	1	1	5	6	138	1
3	5	4	4	1	3	4	2	38	1	3	6	8	72	1	3	1	3	105	0	3	4	2	139	1
2	5	4	4	1	2	1	5	38	0	2	2	4	72	1	2	3	2	105	1	2	1	6	139	0
1	2	3	4	1	1	4	6	38	1	1	3	4	72	1	1	6	2	105	1	1	4	6	139	1
3	5	4	5	1	3	1	8	39	1	3	5	4	73	1	3	1	4	106	0	3	6	8	140	1
2	5	4	5	1	2	3	2	39	1	2	5	4	73	1	2	3	5	106	0	2	2	4	140	1
1	2	3	5	1	1	6	2	39	1	1	2	4	73	0	1	6	2	106	1	1	3	5	140	0
3	1	3	6	0	3	6	8	40	1	3	5	4	74	1	3	3	4	107	1	3	6	8	141	1
2	3	5	6	0	2	2	2	40	0	2	5	4	74	1	2	4	8	107	1	2	2	4	141	1
1	6	2	6	1	1	3	4	40	1	1	2	8	74	0	1	5	6	107	1	1	3	5	141	0
3	3	5	7	0	3	5	4	41	1	3	5	4	75	1	3	1	8	108	1	3	5	4	142	1
2	4	8	7	1	2	5	4	41	1	2	5	4	75	1	2	3	2	108	1	2	5	4	142	1
1	5	6	7	1	1	2	5	41	0	1	2	4	75	0	1	6	2	108	1	1	2	4	142	0
3	3	4	8	1	3	6	8	42	1	3	5	4	76	1	3	5		109		3	5	4	143	1
2	4	8	8	1	2	2	4	42	1	2	5	4	76	1	2	5	4	109	1	2	5	4	143	1
1	5	6	8	1	1	3	4	42	1	1	2	3	76	1	1	2	3	109	1	1	2		143	
3	3	4	9	1	3	6	8	43	1	3	5	4	77	1	3	5	4	110	1	3	5	4	144	1
2	4	8	9	1	2	2	4	43	1	2	5	4	77	1	2	5	4	110	1	2	5	4	144	1
1	5	6	9	1	1	3	4	43	1	1	2	3	77	1	1	2	3	110	1	1	2	4	144	0
3	1	8	10	1	3	5	4	44	1	3	5	4	78	1	3	5	4	111	1	3	5	4	145	1
2	3	5	10	0	2	5	4	44	1	2	5	4	78	1	2	5	4	111	1	2	5	4	145	1
1	6	2	10	1	1	2	4	44	0	1	2	3	78	1	1	2	3	111	1	1	2	3	145	1
3	1	8	11	1	3	5	4	45	1	3	5	4	79	1	3	5	4	112	1	3	1	8	146	1
2	3	2	11	1	2	5	4	45	1	2	5	4	79	1	2	5	4	112	1	2	3	2	146	1
1	6	2	11	1	1	2	3	45	1	1	2	8	79	0	1	2	4	112	0	1	6	2	146	1
3	3	4	12	1	3	5	4	46	1	3	6	8	80	1	3	2	5	113	0	3	1	8	147	1
2	4	8	12	1	2	5	4	46	1	2	2	4	80	1	2	6	3	113	1	2	3	2	147	1
1	5	6	12	1	1	2	3	46	1	1	3	4	80	1	1	1		113		1	6	2	147	1
3	3	4	13	1	3	6	8	47	1	3	6	8	81	1	3	3	4	114	1	3	2	3	148	1
2	4	8	13	1	2	2	4	47	1	2	2	4	81	1	2	4	8	114	1	2	6	3	148	1
1	5	6	13	1	1	3	4	47	1	1	3	4	81	1	1	5	6	114	1	1	1	3	148	0
3	5	4	14	1	3	4	2	48	1	3	4	2	82	1	3	3	4	115	1	3	2	3	149	1
2	5	4	14	1	2	1	1	48	1	2	1	1	82	1	2	4	8	115	1	2	6	3	149	1
1	2	4	14	0	1	4	6	48	1	1	4	6	82	1	1	5	6	115	1	1	1	1	149	1
3	6	8	15	1	3	4	2	49	1	3	4	2	83	1	3	3	4	116	1	3	2	3	150	1
2	2	4	15	1	2	1	6	49	0	2	1	5	83	0	2	4	8	116	1	2	6	3	150	1
1	3	4	15	1	1	4	6	49	1	1	4	6	83	1	1	5	6	116	1	1	1	3	150	0
3	5	4	16	1	3	4	2	50	1	3	6	8	84	1	3	1	1	117	0	3	2	3	151	1
2	5	4	16	1	2	1	6	50	0	2	2	4	84	1	2	3	2	117	1	2	6	3	151	1
1	2	4	16	0	1	4	6	50	1	1	3	4	84	1	1	6	2	117	1	1	1	3	151	0
3	6	8	17	1	3	4	3	51	0	3	3	5	85	0	3	3	4	118	1	3	2	3	152	1
2	2	4	17	1	2	1	5	51	0	2	4	8	85	1	2	4	8	118	1	2	6	3	152	1
1	3	4	17	1	1	4	6	51	1	1	5	6	85	1	1	5	6	118	1	1	1	1	152	1
3	6	8	18	1	3	4	2	52	1	3	3	4	86	1	3	3	4	119	1	3	2	3	153	1
2	2	4	18	1	2	1	5	52	0	2	4	8	86	1	2	4	8	119	1	2	6	3	153	1
1	3	4	18	1	1	4	6	52	1	1	5	6	86	1	1	5	6	119	1	1	1	1	153	1
3	1	8	19	1	3	2	3	53	1	3	4	2	87	1	3	1	8	120	1	3	2	3	154	1
2	3	5	19	0	2	6	3	53	1	2	1	6	87	0	2	3	2	120	1	2	6	3	154	1
1	6	2	19	1	1	1	3	53	0	1	4	6	87	1	1	6	2	120	1	1	1	1	154	1
3	4	2	20	1	3	4	2	54	1	3	1	8	88	1	3	3	4	121	1	3	2	3	155	1
2	1	1	20	1	2	1	1	54	1	2	3	2	88	1	2	4	8	121	1	2	6	3	155	1
1	4	6	20	1	1	4	6	54	1	1	6	2	88	1	1	5	6	121	1	1	1	1	155	1
3	4	2	21	1	3	4	2	55	1	3	3	5	89	0	3	3	4	122	1	3	2	3	156	1
2	1	1	21	1	2	1	1	55	1	2	4	8	89	1	2	4	8	122	1	2	6	3	156	1
1	4	6	21	1	1	4	6	55	1	1	5	6	89	1	1	5	6	122	1	1	1	3	156	0
3	6	8	22	1	3	4	2	56	1	3	4	2	90	1	3	3	4	123	1	3	2	3	157	1
2	2	4	22	1	2	1	5	56	0	2	1	1	90	1	2	4	8	123	1	2	6	3	157	1
1	3	5	22	0	1	4	6	56	1	1	4	6	90	1	1	5	6	123	1	1	1	1	157	1
3	4	2	23	1	3	6	8	57	1	3	4	2	91	1	3	3	4	124	1	3	1	3	158	0
2	1	1	23	1	2	2	4	57	1	2	1	1	91	1	2	4	7	124	0	2	3	5	158	0
1	4	6	23	1	1	3	4	57	1	1	4	7	91	0	1	5	6	124	1	1	6	2	158	1
3	1	3	24	0	3	4	2	58	1	3	6	8	92	1	3	4	2	125	1	3	6	8	159	1
2	3	2	24	1	2	1	1	58	1	2	2	4	92	1	2	1	1	125	1	2	2	4	159	1
1	6	2	24	1	1	4	6	58	1	1	3	4	92	1	1	4	6	125	1	1	3	3	159	0
3	1	3	25	0	3	5	4	59	1	3	1	3	93	0	3	3	4	126	1	3	6	8	160	1
2	3	2	25	1	2			59		2	3	5	93	0	2	4	8	126	1	2	2	4	160	1
1	6	2	25	1	1	2	5	59	0	1	6	2	93	1	1	5	6	126	1	1	3	4	160	1
3	4	2	26	1	3	4	2	60	1	3	4	2	94	1	3	5	4	127	1	3	3	8	161	0
2	1	1	26	1	2	1	5	60	0	2	1	1	94	1	2	5	4	127	1	2	4	8	161	1
1	4	6	26	1	1	4	6	60	1	1	4	6	94	1	1	2	3	127	1	1	5	6	161	1
3	3	8	27	0	3	4	2	61	1	3	1	3	95	0	3	6	8	128	1	3	3	4	162	1
2	4	8	27	1	2	1	6	61	0	2	3	5	95	0	2	2	4	128	1	2	4	8	162	1
1	5	6	27	1	1	4	6	61	1	1	6	2	95	1	1	3	4	128	1	1	5	6	162	1
3	1	8	28	1	3	4	2	61	1	3	4	2	96	1	3	1	8	129	1	3	1	8	163	1
2	3	2	28	1	2	1	1	62	1	2	1	1	96	1	2	3	2	129	1	2	3	5	163	0
1	6	2	28	1	1	4	6	62	1	1	4	7	96	0	1	6	2	129	1	1	6	2	163	1
3	4	2	29	1	3	4	2	63	1	3	6	8	97	1	3	1	8	130	1	3	1	8	164	1
2	1	5	29	0	2	1	5	63	0	2	2	4	97	1	2	3	2	130	1	2	3	5	164	0
1	4	6	29	1	1	4	6	63	1	1	3	2	97	0	1	6	2	130	1	1	6	2	164	1
3	4	2	30	1	3	5	4	64	1	3	3	4	98	1	3	1	8	131	1	3	1	8	165	1
2	1	1	30	1	2	2	4	64	1	2	4	8	98	1	2	3	5	131	0	2	3	5	165	0
1	4	8	30	1	1	2	4	64	0	1	5	6	98	1	1	6	2	131	1	1	6	2	165	1
3	5	4	31	1	3	2	3	65	1	3	3	4	99	1	3	2	7	132	0	3	1	3	166	0
2	5	4	31	1	2	6	3	65	1	2	4	8	99	1	2	6	3	132	1	2	3	5	166	0
1	2	5	31	0	1	1	5	65	0	1	5	6	99	1	1	1	1	132	1	1	6	2	166	1
3	2	3	32	1	3	2	3	66	1	3	1	3	100	0	3	2	3	133	1	3	1	8	167	1
2	6	3	32	1	2	6	3	66	1	2	3	5	100	0	2	6	3	133	1	2	3	6	167	0
1	1	1	32	1	1	1	3	66	0	1	6	2	100	1	1	1	7	133	0	1	6	2	167	1
3	2	3	33	1	3	2	3	67	1	3	3	4	101	1	3	6	8	134	1	3	5	4	168	1
2	6	3	33	1	2	6	3	67	1	2	4	8	101	1	2	2	4	134	1	2	5	4	168	1

Study 7: Simulated identification experiment

Simult. = simultaneous; Line1_x (and likewise, Line2_x,Line3_x) = accuracy of identification outcome of lineup 1. fa = false alarm, ir = incorrect rejection; cr = correct rejection; h = hit; do = foil identification.

Subj	Array	Structure	Lineup1	Lineup2	Lineup3	Line1_x	Line2_x	Line3_X	Subj	Array	Structure	Lineup1	Lineup2	Lineup3	Line1_x	Line2_x	Line3_X
1	c	sequential	4	4	0	fa	h	cr	35	d	sequential	0	0	0	cr	lr	cr
2	d	simult	3	7	0	fa	h	ir	36	d	sequential	3	0	6	fa	cr	h
3	c	simult	0	0	5	ir	cr	h	37	d	sequential	2	0	6	fo	cr	h
4	c	sequential	0	2	0	ir	fa	ir	38	d	simult	3	0	0	fo	cr	cr
5	d	simult	6	0	6	h	cr	h	39	d	sequential	8	3	6	fa	fa	h
6	d	simult	1	0	6	fa	cr	h	40	c	sequential	3	0	5	h	cr	h
7	c	simult	0	4	0	cr	h	cr	41	d	sequential	2	0	0	fa	fr	cr
8	c	sequential	1	0	5	fa	cr	h	42	c	sequential	3	4	0	h	h	cr
9	c	simult	0	0	0	ir	ir	cr	43	d	sequential	0	7	0	cr	h	cr
10	d	sequential	2	0	0	fa	ir	cr	44	d	sequential	0	0	6	cr	cr	h
11	c	sequential	0	4	0	cr	h	cr	45	d	simult	2	3	2	fa	fa	fa
12	d	sequential	0	3	6	cr	fa	fo	46	c	simult	5	0	6	fa	ir	fa
13	d	simult	0	0	4	cr	ir	fa	47	c	simult	1	0	5	h	cr	h
14	c	simult	8	0	2	fa	ir	fa	48	d	simult	6	2	6	h	fa	h
15	c	sequential	7	1	0	fo	fo	cr	49	d	sequential	0	7	0	cr	h	cr
16	d	simult	3	0	6	fo	cr	h	50	c	sequential	3	0	0	h	fr	cr
17	c	simult	3	5	0	fa	fo	cr	51	c	sequential	0	1	5	fr	fa	h
18	d	sequential	2	6	8	fo	fa	fo	52	d	sequential	6	0	0	h	fr	cr
19	c	simult	1	0	0	h	cr	cr	53	d	simult	3	3	6	fa	fo	h
20	d	simult	7	0	6	fa	ir	h	54	c	sequential	0	0	0	cr	ir	cr
21	c	sequential	3	0	4	h	ir	fa	55	d	sequential	0	7	0	cr	h	cr
22	c	sequential	3	4	3	h	h	fa	56	d	simult	0	2	0	ir	fa	cr
23	d	simult	3	1	6	fa	fa	fa	57	c	sequential	8	2	5	fa	fa	h
24	c	sequential	0	0	0	cr	ir	cr	58	c	simult	0	6	0	ir	fa	cr
25	d	sequential	0	0	0	ir	ir	cr	59	c	simult	3	4	7	fa	h	fa
26	d	sequential	2	0	6	fo	cr	h	60	d	sequential	6	0	6	h	cr	h
27	d	sequential	6	7	8	h	h	fa	61	c	sequential	0	0	0	ir	ir	cr
28	c	simult	6	0	0	fa	ir	cr	62	c	simult	0	1	5		fa	h
29	d	sequential	6	3	0	h	fa	ir	63	c	simult	0	0	0		ir	ir
30	d	simult	3	0	0	fo	cr	cr	64	c	simult	0	5	5		fo	h
31	d	sequential	6	0	0	h	ir	cr	65	c	simult	0	0	5		cr	h
32	c	simult	6	0	0	fa	ir	cr	66	d	simult	0	7	0	cr	h	cr
33	c	simult	3	5	5	fa	fa	h	67	d	simult	6	0	6	h	cr	h
34	c	simult		4	4		h	fa	68	d	simult	7	0	8	fa	ir	fa