Chapter

How Many Faces Can We Remember? Why This Matters when Assessing Eyewitnesses

Alicia Nortje[1,*], Colin Tredoux[1,2] and Annelies Vredeveldt[3]

[1]Department of Psychology, University of Cape Town, Cape Town, South Africa
[2]Université de Toulouse, Jean Jaurés, Toulouse, France
[3]Department of Criminal Law and Criminology, Vrije Universiteit Amsterdam, Amsterdam,The Netherlands

Abstract

Most research on eyewitness memory has focused on single-perpetrator crimes. However, crimes to which eyewitnesses may bear testimony are often committed by groups of perpetrators. A consequence of researching only single-perpetrator crimes is that we know very little about how set size (i.e. the number of faces) at encoding impacts recognition performance. We do not know much more about this question in the face recognition literature either but the small extant literature does appear to converge on one conclusion, namely that recognition performance is worse for larger set sizes. In the case of eyewitness memory, the presence of multiple perpetrators poses an additional unique question: Eyewitnesses not only need to identify perpetrators, but also need to testify to the perpetrators' actions. Few researchers have investigated this second aspect. In this chapter, we review literature in the areas of face recognition and eyewitness memory to shed light on these questions, and present two laboratory studies that test the effects of set size on face and person recognition. Results show that recognition performance decreases as a function of set size, but that this is differentially true for faces, and roles, and is in fact dramatically reduced when faces and roles are paired. There are serious applied implications for this latter finding in particular.

[*] Corresponding Author address: Alicia Nortje, Department of Psychology, University of Cape Town, Rondebosch 7701, South Africa. Email: Alicia.nortje@gmail.com.

Keywords: multiple-perpetrator crimes, set size, eyewitness memory, paired memory, load at encoding

INTRODUCTION

There is little doubt that human beings have the ability to recognize a great many faces, certainly in the thousands. This ability continues into old age, even for faces that we last encountered many decades earlier (Bahrick, Bahrick, & Wittlinger, 1975). However, most of our knowledge about face recognition capacity derives from studies of familiar face recognition – that is, recognition of faces of people with whom we are well acquainted and have likely encountered many times. We know much less about our capacity to recognize relatively unfamiliar faces that we have encountered only briefly. This distinction is important. It has a direct bearing on applied questions, specifically, the reliability of eyewitness identifications, since eyewitnesses are almost always tasked with identifying people they have encountered only once. The advent in the 1980s of highly accurate DNA matching techniques for assessing circumstantial evidence of identity has dramatically demonstrated the potential unreliability of eyewitness identifications. At the time of writing, DNA matching has resulted in the exoneration of over 350 prisoners in the United States since 1989, and erroneous eyewitness identification played a major role in over 70% of these cases (Innocence Project, 2017). Psychologists have been researching the basis for eyewitness errors, and some of this research has been important in the revision of aspects of the US criminal justice system (e.g., Wells, 2006, but also Clark, 2012).

But it is important to point out that research on eyewitness identification is driven by research questions that focus almost exclusively on single-perpetrator crimes, in other words, an identification following a simulated crime that involves encoding and recognizing a single face. We know little about how eyewitnesses perform after witnessing a crime committed by more than one perpetrator. Does eyewitness recognition performance decline with increasing numbers of perpetrators? If so, what form does this forgetting function take? The near-exclusive focus on single perpetrator crimes is problematic, since many criminal incidents are committed by groups of two or more perpetrators. For example, it is estimated that roughly 20% of violent crime in the U.S. (Sourcebook of Criminal Justice Statistics, 2008) and between 46% and 70% of crimes against minorities in the EU are committed by groups of two or more perpetrators (European Union Agency for Fundamental Rights, 2012). Research shows that between 10% and 33% of reported rapes in the U.S. (Franklin, 2004), 23% of sexual assaults in Australia (Australian Bureau of Statistics, 2004), 19% of rapes in Southwark in the United Kingdom (Curran & Millie, 2003) and between 30% (Maw, 2012) and 50% of rape cases in South Africa (Swart, Gilchrist, Butchart, Seedat, & Martin, 2000) are committed by multiple perpetrators. This lacuna *could* be addressed in part by referring to basic research on face recognition processes, particularly the effect of (face) set size on recognition memory, if not for the paucity of research on this topic itself. We will discuss a few relevant studies in this chapter.

There is an additional problem when multiple perpetrators are involved, which is not obvious in single perpetrator scenarios nor usually addressed in basic or applied research. Eyewitness identification following a multiple-perpetrator crime is a two-stage process: Firstly, the lineup tests the ability of the eyewitness to recognize the perpetrator (if he/she is present, and to reject all the known incorrect options in the lineup). This is *followed* by a test in which the eyewitness must recall the acts the perpetrator committed during the crime. In a single-perpetrator scenario, an eyewitness identification of the perpetrator *implies* that the identified person committed the acts in question. However, this implication is not automatically present when an eyewitness identifies someone in a lineup following a multiple-perpetrator crime, since the perpetrators may have committed different actions. Therefore, the second stage has to be tested explicitly, requiring the witness to state what action(s) the identified perpetrator committed (personal communication, Captain K. Speed, South African Police Services, 18 September 2015). This important question has been neglected in the literature, even though it has far-reaching consequences for the criminal justice system, including i) assisting the police investigation, ii) building a prosecutable case, and iii) passing an appropriate sentence.

The aim of this chapter is to review what is known about these questions – that is, how set size, or number of perpetrators in a crime scenario, affects facial recognition and eyewitness memory, and how it affects testimony about the event itself as a consequence. There is as yet no published data on the crucial question of the effect of set size on role-person matching, but we will present some new data that directly tests this effect.

THE EFFECT OF SET SIZE ON MEMORY RETENTION FOR NON-FACE IMAGES

We start by considering the small literature on the impact of set size on memory for non-face visual images. Performance on non-face visual stimulus material is a useful baseline reference even though faces are inherently similar to one another in overall appearance (far more so than most other sets of objects), and this may make recognition more difficult. It is important to know whether limits for face and non-face material are similar as a form of benchmark.

People are able to retain large numbers of non-face images in memory. In fact, when tested on 200 non-face images in a 1965 study, participants were able to recognize roughly 95% of these immediately afterwards (Nickerson, 1965), and even after a one year delay recognized 30% of those items (Nickerson, 1968). These results are similar to those reported by Shepard (1967), who found that participants were able to recognize 97% of 612 images after a brief delay. Standing (1973) studied set size at encoding, as well as the 'vividness' of images. His participants studied items in sets varying between 20 and 10,000 in size, which they had to recognize two days later. Participants studied one of three types of items comprising words, 'vivid images', or 'normal images' (i.e. non-vivid, and not words). The words were randomly selected from an English dictionary, and vivid images were images with highly distinctive features (e.g., a dog smoking a pipe). Memory for both types of image (vivid, normal) was superior to that for words, and this difference became more pronounced as set size increased. Estimates of memory storage were impressive, especially for large set sizes: Participants were able to recognize about 88% and 77% of 1000 vivid and normal images respectively, after a brief delay, and even when set size was increased to 4,000 and 10,000 images recognition was still high, at 62% and 66% respectively.[1]

MEMORIAL LIMITS FOR FACES

Does our large memory capacity for non-face visual images mean that we also have a large capacity for human faces? A strong form of evidence would be from systematic manipulation of set size at encoding, and measurement of its effect on recognition. However, studies that have reported such manipulations are often hard to interpret. Most face recognition studies that have included large set sizes at encoding have used within-participants designs. In fact, it is an accepted method to include many trials within such an experiment to control for characteristics of the stimuli (e.g., distinctiveness, attractiveness) that may affect recognition performance, and often these trials will vary dependent on whether the target is absent or present. In contrast, most eyewitness studies do not control for these characteristics, and have stimulus confound problems due to using only single-target encoding (e.g., participants are tested on one target face, which may be distinctive and which does not vary between participants). The authors of a large meta-analysis of the face recognition literature highlight this problem – the difficulty of interpreting and comparing studies from these two research areas - by stating that their results are confounded by methodological problems in and across these two research areas (Shapiro & Penrod, 1986).

[1] One caveat of this research is that memory for all items, especially the much larger set sizes, .i.e. 10,000 images, was not tested. Recognition memory was tested on a sample of items from the original encoding set, and then memorial limits were estimated using the following formula:

Estimated memory capacity = S(T-2E)/T

The symbols, S, E, and T refer to the set size at encoding, the average number of errors at recognition, and the number of recognition test trials respectively. However, what is assumed here is that the rate of guessing (which is double the number of errors [2E]) is *constant*, and thus memory capacity is linear across set sizes. It is possible that participants' response patterns may change as set size increases, so that performance drops in a non-linear way. Thus this estimated memory capacity may be inflated.

Table 1. Characteristics and results from face recognition studies that have manipulated set size at encoding

		Load (number of faces)		Results		
Study	Materials	Encoding	Recognition	d’	Hit rate	FA rate
Podd, 1990	Photofit faces	20	40	1.55 (n/a)	0.74 (n/a)	0.20 (n/a)
		35	70	1.32 (n/a)	0.68 (n/a)	0.23 (n/a)
		50	100	1.10 (n/a)	0.65 (n/a)	0.24 (n/a)
Metzger, 2002	ComPhotoFit faces	10	20	2.05 (0.89)	0.79 (0.13)	0.17 (0.12)
		20	40	1.54 (0.72)	0.72 (0.11)	0.20 (0.10)
		30	60	1.23 (0.58)	0.67 (0.11)	0.25 (0.12)
Lamont et al., 2005	Colour photographs	20	40	2.07 (0.96)	0.78 (0.17)	0.19 (0.14)
		40	80	1.75 (0.76)	0.78 (0.13)	0.22 (0.15)

Note: d’ is a measure of discriminability; FA rate = false alarm rate; Rates are proportions, and standard deviations are reported in parentheses.

However, there are some face recognition studies that have explicitly manipulated set size at encoding, and in which the effect of set size seems clear. Podd (1990) was one of the first to explicitly test the effect of memory load (or set size) at encoding by manipulating the number of faces[2] to be encoded (20, 35 or 50). Recognition performance decreased as set size increased, which was due to a decrease in hit rate specifically (for more detail see Table 1). Podd argued that higher set sizes could decrease performance because of a greater load at encoding, or a greater load at recognition. Thus, the effect may not only be due to the larger set size at encoding, but could result from the larger number of items at test – an inevitable consequence of larger encoding set sizes. Load at recognition does impact on performance (Shapiro & Penrod, 1986), and it is possible that performance decreases as the length of the test sequence increases. Podd therefore employed an interesting method to control for load at recognition and the negative impact of sequence length, by keeping the first forty items (20 old, 20 new) at recognition constant across the various experimental groups. This allowed him to compare the effect of set size at encoding on recognition performance *independently* of set size at recognition. The results showed that there was a significant difference in discriminability (i.e., the ability to accurately recognize old faces and reject new faces, see Macmillan & Creelman, 2004) across set sizes. Participants performed increasingly worse as set size at encoding increased, primarily due to a decrease in hits. Analysis of the remaining recognition trials (i.e. those after the first 40 images) showed little effect of sequence length: there was very little difference between performance on the first 40 trials and the

[2] Note that these were ‘Photofit’ faces, rather than faces of real people.

remaining trials. In another face recognition study, Metzger (2002) studied groups of children, students and adults, who were randomly assigned to three encoding groups (10, 20 or 30 composite face images). He reported three statistics: hits, false alarms and *d'* (a standard measure of discriminability). The results were clear: *d'* and hits were highest when the set size was low (10 images), but were comparable between set sizes of 20 and 30. However, unlike Podd (1990), Metzger also found evidence of an increased false alarm rate as a function of the increase in set size, especially with 30 face images.

Unlike Metzger (2002) and Podd (1990), who used composite face images generated by computer software, Lamont et al. (2005) tested the effect of set size with face images of real people. They showed either a larger (40) or smaller (20) number of young and old face images to young and old participants. There was a main effect of set size, whereby *d'* decreased following encoding of the larger number of faces (*d'* = 2.07 versus *d'* = 1.75, Cohen's *d* = 0.37*)*. This difference in discriminability was due to a significant increase in the false alarm rate, whereas the hit rate was unaffected by memory load. These results corroborated what seems to be a reasonable expectation, namely that recognition performance is better for lower set sizes, and this recognition performance appears to decrease as set size increases (although we cannot currently assess whether it does so linearly). But unlike Podd (1990), who found a decrease in discriminability due to a decrease in the hit rate and Metzger (2002) who found a decrease in discriminability due to changes in the hit rate and false alarm rate, Lamont and colleagues (2005) found a decrease in discriminability due to an increase in the false alarm rate only.

One explanation for these discrepancies may be the manner in which Podd (1990) coded hit rate and false alarm rate. Instead of making a binary response (Yes/No, or Old/New), participants responded 1, 2, 3 or 4, which meant "very sure old", "fairly sure old", "fairly sure new", and "very sure new" respectively. Hit rates comprised 1 and 2 responses (i.e. "very sure old" and "fairly sure old") for old faces, whereas false alarm rate comprised 1 and 2 (i.e. "very sure old" and '"fairly sure old") for new faces. Another explanation could be that both Metzger (2002) and Lamont and colleagues (2005) did not employ any delay between encoding and recognition, whereas Podd (1990) manipulated delay so that it was either 10 minutes, one week or two weeks. It is possible that hit rate is more sensitive to delay than false alarm rate: Perhaps delay reduces the strength of the memory trace, thus causing participants to reject old (i.e. previously seen) faces. It is also worthwhile to note again that both Podd (1990) and Metzger (2002) used artificial faces/face composites, whereas Lamont and colleagues used real faces. While we are not able to compare the results directly or statistically, both Podd and Metzger reported what appears to be comparable hit rates at set size 20, and set sizes 30 and 35, whereas Lamont and colleagues reported higher hit rates at set size 20. This could be due to stimulus difficulty – perhaps composite faces are more difficult to discriminate?

The detrimental presence of a second target face is not limited to studies on memory but is also present in visual search tasks. Mestry, Menneer, Cave, Godwin and Donnelly

(2017) demonstrated that participants were slower and less accurate when they had to search for two target faces than when they had to search for only one. This effect was replicated across three studies that manipulated similarity between the target faces and foils. Interestingly, participants showed a tendency to prefer one target over another in the dual face condition – that is, participants would perform better and faster at recognizing one of the two faces rather than performing equally well at both. This finding is in line with the eyewitness study by Wells and Pozzulo (2006), where participants better recognized and described one perpetrator (i.e., the accomplice) over another (i.e., the main assailant) in a staged crime. However, in that experiment, recognition performance for the main assailant was at chance level (.16), whereas participants in Mestry and colleagues' study (2017) performed better than chance when searching for both the preferred and non-preferred target faces in Experiments 1 and 3 in their article (Experiment 1: 0.79 versus 0.63; Experiment 3: 0.95 versus 0.78; Chance = 0.5). As mentioned previously, eyewitness studies (like Wells & Pozzulo, 2006) and face recognition studies use different methods that may impact the results, and the study by Mestry and colleagues that is reported here is a search or delayed matching task. The difference in performance could be due to properties of the target faces (for example, distinctiveness or criminality), or of the lineup (such as bias, or effective size, see Malpass, Tredoux, & McQuiston-Surrett, 2007), or capacity limits of perception and memory. Mestry and colleagues (2017) controlled for properties of the target images and the recognition arrays by using different stimuli for each of the three experiments, and employed a repeated-measures design with roughly 250 trials. Thus, their findings may be due to capacity limits of visual working memory rather than the former explanaions, such as properties of the face, or lineup, or experimental artefact.

There are thus only a few studies that have a direct bearing on the question of set size and its effect on face recognition, but they do show the expected negative effect of increasing load as seen in the face recognition literature. It is not clear whether increasing set size drives down recognition performance through decreasing hits or increasing false alarms. Most pertinent for present concerns may be the absence of information in these studies about memory for face-connected attributes, such as roles, attributes, or actions.

For this reason, we ran an experiment that tested the effects of set size on face and role / attribute recognition, with 70 participants from the University of Cape Town (Nortje, Tredoux, & Vredeveldt, 2015). Participants studied either one, two, three, five, ten, fifteen or thirty faces, and corresponding attributes, on which they were subsequently tested following a negligible delay (roughly 90 seconds). Each attribute was one sentence, with a maximum of 10 words - for example "He makes his own beer" or "He hates raisins". Each face was shown alone for three seconds, then accompanied by an attribute that appeared below it for three seconds, and then the attribute appeared alone for three seconds. Following a distractor task, participants completed three types of tests: face recognition, attribute recognition, and face-attribute pairing.

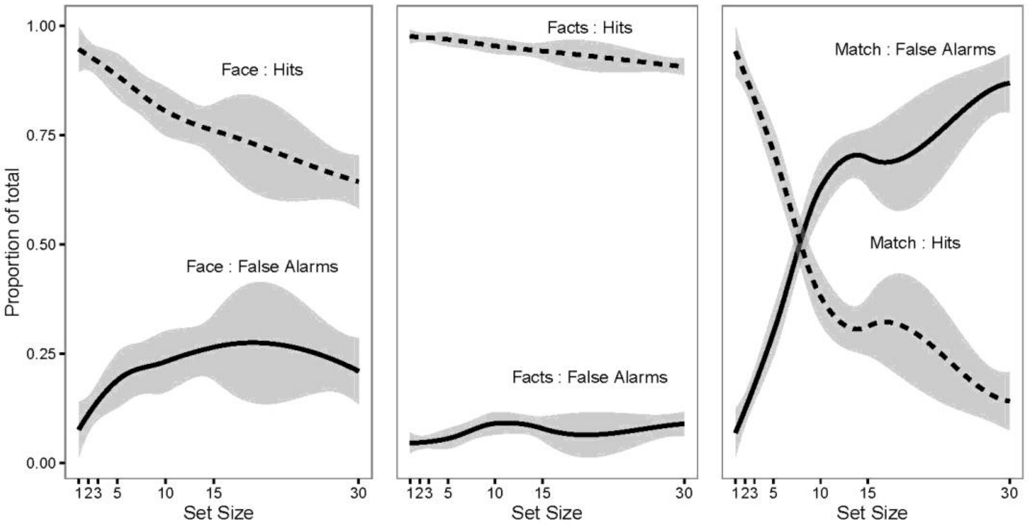

Figure 1. Hits and false alarms as a function of Set Size, for three types of encoding material. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, with span = 1.3. In these graphs, ‘Facts’ refers to the attribute, and ‘Match’ refers to the face-attribute pairing.

Figure 1 shows the hit and false alarm rate for faces, attributes, and pairings across set size. For present purposes, the most interesting result was that there was a significant interaction between Set Size and the Type of Test, for both hit rate, $F(10.528, 110.539) = 46.46$, $p < .001$, $\eta_p^2 = .816$, and false alarm rate, $F(12, 126) = 40.685$, $p < .001$, $\eta_p^2 = .795$. The nature of these interactions can be understood by examining Figure 1. The proportion of hits decreases across Set Size for all three Types of Test, but the rate of decrease is much greater for face-attribute pairing than for the two other Types of Test. A similar pattern (but in the opposite direction) is true for false alarms: the proportion of false alarms increase across Set Size for all three types of test, but moreso for the face-attribute pairing.

EYEWITNESS STUDIES: MANIPULATING SET SIZE AT ENCODING

An important applied domain of the theoretical question about the effect of set size is eyewitness memory for multiple-perpetrator crimes. In such an event, a witness observes a crime that is committed by more than one person (i.e. a perpetrator). This is a complex visual scenario where each perpetrator may perform a different action (e.g., drawing a weapon) and voice a different command (e.g., demand a bag, or threaten the victim). It is vital that eyewitnesses are able to recall the event successfully in their statement to the police, which may in turn lead to their participating in an identification parade. Until recently, most eyewitness studies assessed memory for staged crimes enacted by a single perpetrator. There is now a growing interest in multiple perpetrator crimes. Despite this interest, there are fewer than 15 articles on this topic (see Table 2). In the bulk of these studies, recognition memory was tested with an identification task, and role recognition was tested in only one study.

Table 2. Results of studies that tested eyewitness memory for multiple perpetrators (with lineups)

		Load					
Study	Total sample size[e]	Encoding	Test	Lineups	Lineup Size	Hits	FAs
Egan, Pittner, & Goldstein, 1977	86	2	1	TP	5	0.91	-
Clifford & Hollin, 1981	60	1 [b]	1 MA	TP	10	0.35	-
		3 [b]	1 MA	TP	10	0.3	-
		5 [b]	1 MA	TP	10	0.15	-
Shepard, 1983 (Experiment 4)	n/a	2	1	TP/TA [a c]	9 (1 target)	0.3	0.52
Schiff, Banka, & de Bordes Galdi, 1986	84	6 [b]	6	TP only [a c]	18	0.65	0.35[d]
Jacob, 1994	144	3	3	TP/TA [a c]	6 (sim.) vs 18 (seq.)	0.34	0.13
Vanderwal, 1996	144	3	3	TP/TA [a c]	18 (sim.) vs 6 (seq.)	0.17	0.17
Laldin, 1997	168	2	1	TP/TA [a c]	10 (2 targets)	0.3	0.52
		3	3	TP/TA [a c]	18	0.39	0.13
Megreya & Burton, 2006 (Experiment 1)	20	1	1	TP/TA	10	0.6	0.24
		2	1	TP/TA	10	0.34	0.23
Megreya & Burton, 2006 (Experiment 2)	20	2	1	TP/TA	10	0.37	32.6
Wells & Pozzulo, 2006	150	2 [b]	1 MA	TP/TA [a]	6	0.16	0.41
		2 [b]	1 Accomplice	TP/TA [a]	6	0.28	0.41
Hobson & Wilcock, 2011	72	3 [b]	3	TP/TA [a]	9	0.51	0.41
Megreya & Bindemann, 2011	534	1 [b]	1 MA	TP/TA	10	0.54	0.36
		2 [b]	1 MA	TP/TA	10	0.29	0.38

Note. Role accuracy is not reported in this table and standard deviations are not available. MA denotes 'Main Assailant'. For example, Clifford and Hollin (1981) tested for the main assailant.

[a] These studies tested recognition for all targets that appeared at encoding.

[b] These studies assigned roles to the targets at encoding.

[c] These studies included all the suspects in one lineup, rather than one suspect per lineup.

[d] The authors report 'misses' (i.e. foil identifications in a target-present lineup) as false alarms.

[e] The sample sizes reported here are for the entire study, and are not the experimental cell sizes. It was not possible to reliably estimate the cell sizes as the proportions here have been averaged across experimental conditions.

An early study by Clifford and Hollin (1981) showed that eyewitness recognition accuracy decreases as set size increases. Participants viewed a simulated crime that was committed by one, three, or five perpetrators, and were then tested for their recognition of the main assailant from a ten-person, target-present, simultaneous lineup. Recognition performance in the one and three perpetrator conditions was significantly better than chance (35% and 30% respectively, vs. 16.67%), but performance in the five-perpetrator condition was not (15%).

This pattern is what one would expect, given work reviewed earlier on face and image recognition, but there are some important nuances. Interestingly, recognition performance is compromised even if the (multiple) perpetrators are perceptually and obviously different from one another. Megreya and Bindemann (2011) showed participants a video of a crime committed by either one or two perpetrators, who were either of the same sex (two men or two women) or not (one woman and one man). The authors initially hypothesized that participants who encoded perpetrators of a different sex were less likely to confuse the perpetrators with one another. Their results, however, did not show this: Participants in the single-perpetrator condition outperformed participants in the two-perpetrator condition (54% versus 29% respectively), and the gender of the perpetrators made no difference (two same-sex perpetrators: 28.9% versus two opposite-sex perpetrators: 29.4%). One could argue that this poor performance might be due to divided attention: participants performed poorly because they had to process both perpetrators simultaneously within a limited period of time. However, Megreya and Burton (2006) demonstrated that performance was poor following the encoding of two faces *even* when participants were given unlimited time to study them. This poor performance persisted throughout various experiments – even when participants were warned that they would have to recognize the target face later, and when they knew that the target would be absent from some of the lineups. Overall, recognition performance following encoding of two faces was worse than following encoding of one face only (34% versus 59.5% respectively).

EYEWITNESS STUDIES: DOES LINEUP TYPE IMPACT ACCURACY FOR MULTIPLE PERPETRATORS?

Perhaps the poor performance of eyewitnesses who attempt to identify multiple perpetrators could be attributed to the type of recognition test. For example, eyewitnesses may perform worse when presented with a simultaneous lineup instead of a sequential lineup, or eyewitnesses may perform better if they view one large parade rather than many smaller parades.

Table 3. Procedure and results from three unpublished studies comparing simultaneous and sequential lineup presentations

	Simultaneous			Sequential		
Study	Description	HR	FA	Description	HR	FA
Jacob, 1994	Three simultaneous six-person, TP/TA lineups.	0.47	0.11	One 18-person TP/TA lineup	0.29	0.08
Vanderwal, 1996	One 18-person TP/TA lineup.	0.19	0.07	Three TP/TA lineups	0.14	0.27
Laldin, 1997	Six sets of three photographs, TP/TA presented simultaneous	0.5	0.19	Six sets of three photographs, TP/TA presented sequentially	0.26	0.02

Note. FA rate = False alarm rate; HR are hit rates. These proportions are the averaged hit rate and false alarm rate presented in the three studies. TP and TA lineups consist of a combination of zero, one, two, or three target-suspect combinations.

Three unpublished papers from RCL Lindsay's research laboratory at Queen's University[3] examined the differences between eyewitness identifications using sequential and simultaneous lineups (Jacob, 1994; Laldin, 1997; Vanderwal, 1996; see Table 3). All three studies used the same materials: a 45-second encoding video that depicted a crime committed by three perpetrators who steal a woman's purse, and lineups containing the same foils, innocent suspects, and perpetrators. The overall results were somewhat conflicting: False alarm rate was lower following sequential presentation than the simultaneous presentation of photographs, but hit rate was also lower following a sequential presentation (Jacob, 1994; Laldin, 1997). This is the typical pattern observed for sequential lineups: witnesses make more conservative decisions, which is associated with a decrease in both hits and false alarms (Clark, 2012). However, this pattern was partially-reversed in a third study (Vanderwal, 1996), in which hit rate was again lower for sequential lineups, but false alarm rate was higher.

Wells and Pozzulo (2006) compared three different lineup methods to determine how eyewitness recognition of multiple perpetrators is affected. Their participants watched a video of a staged mugging committed by two perpetrators, who were designated main assailant and accomplice, and then viewed one of three types of lineups for both targets: Two six-person target-present or target-absent simultaneous lineups, two six-person target-present or target-absent sequential lineups, or six two-person lineups consisting of

[3] http://www.queensu.ca/psychology/People/Faculty/Roderick-Lindsay.html.

a pair of foils (or one foil and one target, but never two targets). There was no difference in hit rate among the three lineups formats, but participants were more likely to correctly reject the two-person paired target-absent parades for the accomplice and the main assailant. However, it is difficult to interpret these results without some measure of choosing bias (i.e. *c*, a measure of response bias; Macmillan & Creelman, 2004)

A common practice in multiple perpetrator research is to manipulate the number of perpetrators at encoding, but test recognition memory for only one target (see Table 2). This provides an incomplete picture: Is recognition poor because of the set size, or because of some attribute (e.g., distinctiveness) of the target? That is, participants may have encoded one of the perpetrators who is not presented at test. In fact, only three published studies have tested memory for all the perpetrators in the witnessed event (see Table 2) but the decline in performance with increasing number of perpetrators is remarkable: Only one of 41 participants (Shepherd, 1983), eight of 75 participants (Wells & Pozzulo, 2006), and eight of 72 participants (Hobson & Wilcock, 2011), could accurately identify all perpetrators. In summary, this literature shows that recognition accuracy decreases as set size increases, and this is most striking when participants are tested for *all* the perpetrators, with accuracy levels (i.e., identifying all perpetrators correctly) ranging from 2% to 11%.

EYEWITNESS STUDIES: ROLE IDENTIFICATION

Only one multiple-perpetrator eyewitness study tested for role identification of the perpetrators. Hobson and Wilcock (2011) provided two different types of lineup instructions to their participants prior to viewing the lineup. The first type of instruction was to ‘reflect’ on the role that the perpetrator had performed while making their identification decision. The second type of instruction constituted ‘general’ lineup instructions that did not refer to the role of the perpetrator. After making an identification, all participants were asked to identify the role the perpetrator played in the witnessed event. The results showed that i) the different instructions did not affect identification accuracy, but ii) participants were better overall at recalling the roles of the perpetrators when given the ‘reflection’ instructions. Overall, the average role recall performance for the ‘reflection’ and the ‘general’ instructions groups was 69.3% and 30.3% respectively.

It is still unclear how role identification is impacted by the number of perpetrators. Hobson and Wilcock’s results are promising as they suggest that role identification can be improved through lineup instructions. However, their study did not include a baseline one-perpetrator condition, and set size was not manipulated. Therefore, this study does not provide insight into the impact of set size on role identification.

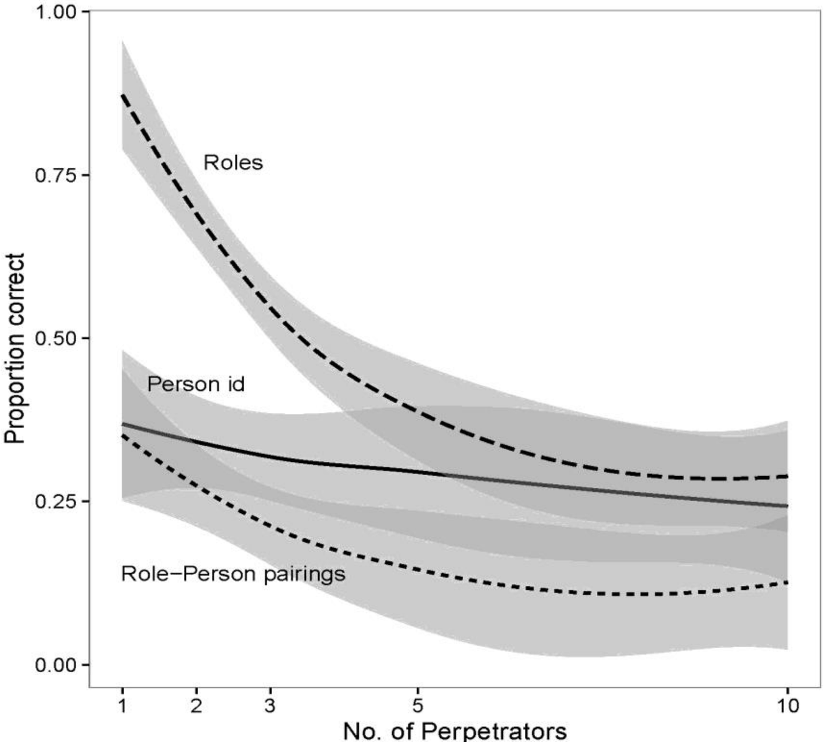

Figure 2. Proportion correct as a function of Set Size, for three types of recognition tests. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, with span = 1.3.

The importance of this question - about how our ability to pair faces and roles is affected by set size – is quite pronounced for applied research such as eyewitness memory, and there is little relevant data (besides the research from Nortje, et al., 2015, and results reported by Hobson and Wilcock, 2011) to help us answer it. As we argued earlier, a unique problem for multiple perpetrator facial recognition, particularly in applied settings, is that the eyewitness' memory must be further tested by asking the witness to declare the role of the person they have identified(e.g., what they did, or what they said). There is some research that suggests that eyewitnesses may find it difficult to do this: Police officers in England, Wales and Northern Ireland report that eyewitnesses who had made multiple identifications would, after making their identification, often provide information about the role of that perpetrator that contradicted the information in their previous statement (Hobson, Wilcock, & Valentine, 2012). It is often not possible for police officers to verify this information, and these disparities could reduce the perceived reliability of an eyewitness (Brewer, Potter, Fisher, Bond, & Luszcz, 1999).

We are able to report some original data here that examines eyewitness memory for faces and roles as a function of set size. We recruited 200 participants from the University of Cape Town, who watched a staged theft in a computer lab committed by either one, two, three, five, or ten perpetrators (Nortje, Deglon, Tredoux, & Vredeveldt, 2016). Following a delay of roughly 30 minutes, participants viewed as many lineups as there were perpetrators and had to make lineup decisions for each. Lineup sequences were a combination of target-present and -absent lineups. Participants made a lineup decision, and if this decision was a positive identification, they then attempted to recall the role performed by the identified target.

The results from this eyewitness study corroborate the findings of the facial recognition study discussed earlier (Nortje, et al., 2015): Recognition performance decreased as the number of perpetrators increased, but decreased especially quickly when trying to link perpetrators to roles (see Figure 2). There was a significant main effect for the Type of Test, $F(1.339, 242.281) = 80.728$, $p < .001$, $\eta_p^2 = .308$. Overall, participants performed better at the Role Recall (*M* = .53, 95% CI [.48, .58], *SD* = .334) than at the Lineup Identifications (*M* = .31, 95% CI [.25, .36], *SD* = .37), but performed better at Lineup Identifications than at correctly recalling the Role for *that* target (i.e. Pairing) (*M* = .21, 95% CI [.16, .26], *SD* = .34). Mean percentage accuracy was significantly different between Identifications and Roles, $F(1, 181) = 51.193$, $p < .001$, $\eta_p^2 = .220$, and between Lineup Identifications and Pairings, $F(1, 181) = 30.615$, $p < .001$, $\eta_p^2 = .145$. There was also a significant interaction between the Type of Test and Set Size, $F(5.354, 242.281) = 5.426$, $p < .001$, $\eta_p^2 = .107$. This interaction can be seen in Figure 2: The proportion of correct responses decreased in all three Tests as Set Size increased, but the rate at which this proportion decreased was different for the three tests. Participants performed quite well at Role identification for small Set Sizes, but this performance dropped precipitously as Set Size increased. Proportion correct for Lineup Identifications and Pairings were similar for small Set Sizes, but Pairings appeared to decrease more quickly than Lineup Identifications as Set Size increased. Thus, for small Set Sizes, participants performed better at Role Recall than at Lineup identifications and Pairings, but performance for all three tests dropped as Set Size increases, with the most notable decrease for Role Recall.

These results corroborate and extend those found in previous research (Nortje, et al., 2015): Memory for faces and actions (or roles) is impaired by an increasing set size, and the pairing of faces to their respective associated actions remains the lowest performance and is most affected by increasing set sizes. These results suggest that eyewitnesses will perform worse at identification tasks if the crime was committed by an increasing number of perpetrators. More concerning is that the ability of eyewitnesses to link criminal actions to perpetrators will be significantly impaired as the number of perpetrators increases, which can have serious repercussions for police investigations, and eyewitness identification.

Conclusion

Our aim in this chapter was to review the evidence on how set size affects person and role recognition, in both the face-recognition and eyewitness-identification literatures. Face recognition researchers have on occasion investigated large set sizes, but have rarely manipulated set size systematically. The few studies that have done this – in both face recognition and eyewitness memory research – show consistent results: Recognition of briefly encountered faces drops considerably as set size increases.

There is as yet no published evidence on the critical question of pairing previously seen faces (or people) with their associated roles or actions, but we reported results from two recently conducted studies that investigated this. The findings from both studies show that memory for previously seen faces is impaired by the number of faces encoded. What our study (Nortje, et al., 2015) shows in particular is concerning, and more so perhaps than the suppression of face recognition memory performance by increasing set size: Participants in our studies showed a particular inability to correctly pair or 'bind' faces they had seen with attributes they had learnt about the faces in question, *despite* being able to recognize these two things independently. This disjunction became increasingly evident with larger set sizes.

This is of considerable concern in practice, due to the prevalence of multiple perpetrator crimes. Eyewitnesses provide testimony about such crimes, and our data suggests that i) these eyewitnesses will show diminished recognition performance as the number of perpetrators increases, and i) they will struggle to accurately recall what the perpetrators did in the crime. This effect is not buttressed by accurate identifications: Even if the eyewitness is able to make a correct identification, this does not necessarily mean that their testimony about *what* that perpetrator did is also accurate. The consequences of inaccurate perpetrator recognition and incorrect testimony about the actions of each perpetrator could impede police investigations, and could lead prosecutors, judges and juries to consider the eyewitness' memory unreliable. Moreover, sentencing is dependent on role recollection: Perpetrators who were less directly involved in the crime may receive reduced sentences. Therefore it is vital that actions are correctly attributed to perpetrators to ensure fair sentences.

In conclusion, the effect of set size on face memory is strong, with serious applied consequences. These consequences appear to be more profound than merely suppressing face recognition memory – eyewitnesses who see multiple perpetrators may be particularly prone to confusing perpetrator identities and roles. Extra caution is urged when eyewitnesses give testimony about events involving multiple perpetrators.

ACKNOWLEDGEMENTS

We would like to extend our thanks to Michaela Deglon, who assisted with data collection. We would also like to thank the actors, who portrayed our ‘perpetrators’, our participants, and the ACSENT laboratory for providing the laboratory space, equipment and resources. Lastly, we would like to thank the National Research Foundation for providing valuable funding for two years.

REFERENCES

Australian Bureau of Statistics. (2004). Sexual assault in Australia: A statistical overview.4523.0. Retrieved from http://www.abs.gov.au/AUSSTATS/ abs@.nsf/ DetailsPage/4523.02004.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty Years of Memory for Names and Faces: A Cross-Sectional Approach. *Journal of Experimental Psychology: General, 104*, 54-75.

Brewer, N., Potter, R., Fisher, R. P., Bond, N., & Luszcz, M. A. (1999). Beliefs and data on the relationship between consistency and accuracy of eyewitness testimony. *Applied Cognitive Psychology, 13*, 297–313. Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238-259.

Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology, 66*, 364-370.

Curran, K., & Millie, A. (2003). Rape and indecent assault: Incidence and service provision in Southwark (Project Report). London, England: Safer Southwark Partnership.

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

European Union Agency for Fundamental Rights (FRA). (2012). Data in Focus Report: Minorities as Victims in Crime (Report No. 6). Retrieved from European Union Agency for Fundamental Rights website: http://fra.europa.eu/en/publication/2012/eu-midis-data-focus-report-6-minorities-victims-crime.

Franklin, K. (2004). Enacting masculinity: Antigay violence and group rape as participatory theater. *Sexuality Research and Social Policy: Journal of NSRC, 1*, 25-40.

Hobson, Z. J., & Wilcock, R. (2011). Eyewitness identification of multiple perpetrators. *International Journal of Police Science & Management, 13*, 286-296.

Hobson, Z., Wilcock, R., & Valentine, T. (2012). Multiple suspect showing: A survey of police identification officers. *Policing, 7*, 79-87.

Innocence Project (2017). Eyewitness Misidentification. Retrieved from https://www.innocenceproject.org/causes/eyewitness-misidentification/.

Jacob, P. (1994). *The feasibility of using multiple-suspect sequential lineups.* (Unpublished Honour's thesis). Queen's University, Canada.

Laldin, S. (1997). *Contextual effects on lineup identification of multiple perpetrators.* (Unpublished Honour's thesis). Queen's University, Canada.

Lamont, A. C., Stewart-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. *Memory & Cognition, 33*, 1017-1024.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. (Second edition). Mahwah, New Jersey: Lawrence Erlbaum Associated, Publishers.

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In *The handbook of eyewitness psychology, Vol II: Memory for people* (pp. 155-178). Lawrence Erlbaum Mahwah, NJ.

Maw, S. (2012). The psychological impact of rape trauma: A longitudinal study of adult female survivors in the Western Cape, South Africa. (Unpublished doctoral thesis). University of Cape Town, Cape Town.

Megreya, A. M., & Bindemann, M. (2011). Identification accuracy for single- and double-perpetrator crimes: Does accomplice gender matter? *British Journal of Psychology. 103*, 439-453.

Megreya, A. M., & Burton, A. M. (2006). Recognising faces seen alone or with others: When two heads are worse than one. *Applied Cognitive Psychology, 20*, 957-972.

Mestry, N., Menneer, T., Cave, K. R., Godwin, H. J., & Donnelly, N. (2017). Dual-target cost in visual search for multiple unfamiliar faces. *Journal of Experimental Psychology: Human Perception and Performance*.

Metzger, M. M. (2002). Stimulus load and age effects in face recognition: A comparison of children and adults. *North American Journal of Psychology, 4*, 51-62.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology, 19*, 155-161.

Nickerson, R. S. (1968). A note on long-term recognition memory for pictorial material. *Psychonomic Science, 11*, 58-58.

Nortje, A., Tredoux, C.G., & Vredeveldt, A. (2015, March). *Remembering multiple faces is harder than you think! The effect of set size on face recognition.* Paper presented at the American Psychology-Law Society conference, San Diego, USA.

Nortje, A, Deglon, M., Tredoux, C**.,** & Vredeveldt, A (2016, November). *Role and offender identification in multiple-perpetrator crimes.* Presentation at the Annual Face Science Seminar, Cape Town, South Africa.

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

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

Schiff, W., Banka, L., & de Bordes Galdi, G. (1986). Recognizing people seen in events via dynamic "mug shots". *The American Journal of Psychology*, *99*, 219-231.

Shepard, R. N. (1967). Recognition Memory for Words, Sentences, and Pictures. *Journal of Verbal Learning and Verbal Behaviour, 6*, 156-163.

Shepherd, J. W. (1983). Identification after long delays. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating witness evidence* (pp. 173-187). Chicester: John Wiley & Sons.

Sourcebook of Criminal Justice Statistics. (2008). Estimated percent distribution of violent victimizations by multiple offenders by type of crime and perceived race of offenders, Retrieved from http://www.albany.edu/sourcebook/pdf/t3312008.pdf.

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology, 25*, 207-222.

Swart, L., Gilchrist, A., Butchart, A., Seedat, M., & Martin, L. (2000). Rape surveillance through district surgeon offices in Johannesburg, 1996-1998: Findings, evaluation and prevention implications. *South African Journal of Psychology, 30*, 1-10.

Vanderwal, A. (1996). *The effects of the sequential presentation of lineups with multiple perpetrators on eyewitness identification.* (Unpublished Honour's thesis). Queen's University, Canada.

Wells, G. L. (2006). Eyewitness identification: Systemic reforms. *Wisconsin Law Review, 2*, 615-644.

Wells, E. C., & Pozzulo, J. D. (2006). Accuracy of eyewitnesses with a two-culprit crime: Testing a new identification procedure. *Psychology, Crime & Law, 12*, 417-427.

RR