Selection of Lineup Foils in Operational Contexts

STEPHEN DARLING1*,†, TIM VALENTINE1 and AMINA MEMON2

1Goldsmiths College, University of London, UK
2University of Aberdeen, UK

SUMMARY

It has been argued (Luus & Wells, 1991) that matching lineup foils to the description of the culprit, rather than the appearance of the suspect, brings benefits to witness performance on identification tasks, in particular by increasing the rate of correct identifications of the culprit. Recently, live identification procedures in the United Kingdom have been replaced by use of video. The reported research investigated whether use of description matching would improve lineup performance when implemented using this new video based system in an ecologically valid experiment. The effect of using moving rather than still video clips was also investigated. Participants witnessed a live staged incident, and attempted to identify the culprit later from police video lineups. Neither use of a description-matching strategy to select foils nor use of moving images produced a reliable improvement in performance for culprit present (CP) or culprit absent (CA) lineups. Copyright © 2007 John Wiley & Sons, Ltd.

Foils play a fundamental role in the construction of lineups. According to Luus and Wells (1991) well chosen foils should have four functions: they allow eyewitnesses to make responses that are known errors, they help control for chance, they help assure that the witness cannot use deductive reasoning and they help to assure that the lineup ‘constitutes a test of recognition memory rather than a test of recall’ (p. 45). Luus and Wells argued that this latter function is better served when foils are selected using a culprit description (CD) strategy rather than a suspect resemblance (SR) strategy.

Under the SR approach, foils are selected on the basis of their visual similarity to the suspect. The aim is to match the suspect as closely as possible to provide a rigorous test of the witness’ ability to recognise the suspect. One criticism of the SR approach is that it is hard to specify the degree of resemblance required. For example, taking the argument that increasing suspect-foil similarity increases lineup fairness to its logical conclusion would result in a lineup of clones (Luus and Wells, 1991), from which identification would be impossible even for a reliable witness. Whilst this latter case is an extension of the argument to absurdity, increasing the similarity of foils should increase the difficulty of the discrimination even for a reliable witness.

In CD matching, foils are chosen to match the witness’ verbal description of the culprit. Beyond this, however, variations in appearance of lineup members are encouraged. The purpose of an identification procedure is to enable a witness to use recognition memory to

*Correspondence to: Stephen Darling, Human Cognitive Neuroscience, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK. E-mail: s.darling@ed.ac.uk
†Present Address: University of Edinburgh, UK.

Copyright © 2007 John Wiley & Sons, Ltd.
access information about the culprit’s appearance that could not be expressed verbally. Therefore, the CD lineup contains only members who cannot be excluded by the witness with reference to their verbal description. Foils that can be excluded on this basis can legitimately appear in SR lineups, but they are unlikely to be selected, and hence would make the lineup less fair. Therefore, Luus and Wells (1991) argue, CD lineups should be less prone to false suspect identifications. Assuming that the verbal description given by a witness is as detailed as possible, and that the foils chosen are all compatible with their descriptions, then the CD strategy may also lead to higher rates of correct identifications. Characteristics that must and must not be present in each particular foil are specified precisely by the witnesses’ description. Beyond this, the lineup should be as heterogeneous as possible. The benefit of such heterogeneity is that the lineup is likely to require less fine grained perceptual discrimination than would be necessary in a SR lineup. Increasing heterogeneity of lineups can increase correct identifications at the cost of also increasing misidentifications (Wells & Lindsay, 1980), because inappropriate foils can be rejected on the basis of the remembered descriptions, thus reducing the effective size of the lineup. The requirement to match to description should negate this problem.

Laboratory studies have provided support for the use of CD foil selection. Wells, Rydell, and Seelau (1993) demonstrated an increased rate of correct identifications but no increase in false identifications with CD lineups. Lindsay, Martin, and Webber (1994) replicated the increase in correct identifications. Juslin, Olsson, and Winman (1996) found a similar pattern to Wells et al. In contrast, two experiments reported by Tunnicliff and Clark (2000) failed to identify any difference in correct or incorrect identifications between CD and SR lineups.

Luus and Wells (1991) draw attention to two important situations where CD lineups would be inappropriate: when the suspect does not match the witness description, and when the witness description is so detailed or idiosyncratic that it would be impossible to locate a sufficient set of distracters. Lindsay et al. (1994) identified the important issue of default values for description matched lineups. These are attributes which are certainly noticed but that are sometimes left out of descriptions, such as age and race. They found that false identification rates were inflated if foils were selected on a description matched basis but were allowed to vary on these default values. However, such default values may be less likely to be omitted by witnesses who have been through the process of making a formal statement to the police.

A further, unexpected, difficulty has recently emerged for description matching: using an innocent suspect as the basis for constructing a resemblance-based lineup can create a circular ‘backfire’ effect where suspect identifications are inflated because the selection of foils is based around that suspect, so that the suspect forms a central tendency in the lineup (Clark & Tunnicliff, 2001; Navon, 1992). Logically, this backfire effect should not apply in the case of description matched lineups. However, there is evidence that this is not the case. Clark (2003) reviewed five studies in which description and resemblance matching were compared. Data showed that for participants who made a choice from a CA lineup, the likelihood of choosing innocent suspects was higher than that predicted by chance, a result that Clark attributed to the backfire effect. Surprisingly, though, this effect was stronger for description matched than resemblance matched lineups (Clark, 2003, p. 636).

Current guidelines in the UK favour the SR approach to foil selection. The current code of practice drawn up under the Police and Criminal Evidence Act (1984), henceforth referred to as PACE, states that ‘The set of images must include the suspect and at least eight other people who, so far as possible, resemble the suspect in age, general appearance
and position in life’ (PACE Codes of Practice, Code D, Annex A, Para. 2, p. 36). In the United States of America the selection of foils for use in lineups was a specific focus of a comprehensive review of procedures (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998) that formed the basis for a US Department Justice guide of best practice for handling eyewitness identification evidence (Technical Working Group on Eyewitness Evidence, 1999). The key recommendation made was that under normal circumstances foils should be selected to match the description of the culprit given by the witness.

Procedures used to establish evidence of eyewitness identification in Britain have recently been modernised. The codes of practice of PACE now indicate that a suspect should initially be offered a video identification in preference to a live identification parade. Two proprietary systems are currently in use in the UK, Video Identification Procedure Electronic Recording (VIPER™) and Profile Matching (PROMAT™). The two systems are very similar in most important respects, and both allow the construction of lineups in accordance with the PACE codes which specify that images used in video identifications must be moving images unless it is impracticable to record a moving image (e.g. when a suspect does not consent to have a moving image taken). PACE code D specifies that any lineup (live or video) must include at least eight foils.

VIPER video lineups from real criminal cases were found to be fairer to the suspects than conventional ‘live’ lineups (Valentine & Heaton, 1999), and VIPER video lineups have been shown to be equally fair to white European and African—Caribbean suspects (Valentine, Harris, Colom Piera, & Darling, 2003), in contrast to live lineups, which have been shown to be less fair for ethnic minorities (Wright & McDaid, 1996).

The aim of the study reported in this paper was to examine the effectiveness of CD matching in the context of operational VIPER lineups as used in the United Kingdom. There are three main differences between the materials used in previous studies of CD matching and those used in VIPER displays. Firstly, VIPER parades are presented sequentially, one after the other, rather than in simultaneously presented arrays. Secondly, a key aspect of the VIPER system (although not a requirement of PACE per se) is the use of a large, cross-referenced and categorically searchable database of stimuli to use as foils. Although this database is large (around 12 000 items) it represents a constraint on foil selection. Finally, typical VIPER images are moving video clips.

The current experiment also included a manipulation of image type: whether the lineup comprised moving video clips or still, full face images. We have recently observed a small effect of image type in CA lineups. A smaller proportion of witnesses who viewed a moving image lineup made an incorrect identification of a foil in CA lineups (Valentine, Darling, & Memon, in press). It was of interest to see if the effect would be observed in the current experiment.

METHOD

Participants

Two hundred participants took part in the experiment. Mean age was 21 years. There were 67 males and 133 females, and all were undergraduate university students. Participation was arranged through an introductory psychology course requirement.

1Note that this sequential mode of presentation is not accompanied by the strict rules advocated by Lindsay, Lea, and Fulford (1991) and Lindsay and Wells (1985).
Design

This experiment was run in two sessions. In Session 1 participants witnessed a staged crime. In Session 2, which was held between 7 and 25 days later (mean = 11 days, SD = 3 days) they were asked to return to attempt to identify the thief from a lineup.

There were three between participants independent variables: Foil selection method (SR or CD), image type (moving or still) and presence of the culprit in the lineup (CP or CA). For the main effects which were the focus of the experimental hypotheses, a priori power analysis using the G-Power 3 computer software (Faul, Erdfelder, Lang, & Buchner, in press) indicated that a sample size of 88 would be appropriate for this analysis in order to observe a medium ($\phi = 0.30$) effect size with a power ($\beta$) of 0.80.

Materials

Four actors played the role of the thief. For each actor, an innocent suspect who closely resembled the culprit was chosen from the VIPER database. A set of eight foils for SR lineups was selected based on resemblance to the suspect. The VIPER database was searched by a number of specific terms based on the appearance of the suspect (age, sex, race, hair colour, hair length, etc.). This produced a subset of samples (usually around 20) from which the eight foils which most closely resembled the suspect were selected by the researcher: this is exactly the same method used by the UK police, where a single identification officer would normally make a similar selection on the basis of suspect resemblance. Eight foils for CD matched lineups were selected individually for each witness (because each witness provided a different description). Descriptions were collected during Session 1 of the experiment, immediately after the witnesses had seen the staged incident and been informed that the incident was part of the experiment. Witnesses were asked to independently write down free recall descriptions of the culprit: instructions emphasized the importance of only including remembered details, and not making guesses. These descriptions were used as the basis of foil selection in CD lineups. CD foils were selected so that they (a) matched the individual witness’ description and (b) varied as much as possible without violating that description. To avoid the problems associated with default values (Lindsay et al., 1994), sex (male), race (white European) and age (16–45) were added to all descriptions if they were not otherwise mentioned. Also, all description items relating to clothing, gait, accent and the height of the culprit were excluded from the descriptor set used to choose foils, as they were irrelevant.

In moving image conditions, each lineup member appeared as a standard VIPER image, that is, a video clip of the person in which they initially faced the camera, then turned to show the right and then the left profile and finally faced the camera again. The image showed the head and upper torso only, and was shot against a uniform grey background. The clip lasted for 15 seconds. In still image conditions, still full face images captured as single frames from the video clips were shown for 15 seconds. Consistency of the culprit images with the foil images was ensured by recording the actors in an active VIPER unit at a local police station, and subjecting the resultant VIPER clips to exactly the same quality control processes that apply to all of the images (suspect or foil) recorded by VIPER.

Session 1 procedure

Participants attended the laboratory in groups (mean size = 5) having signed up to undertake a ‘health and mood questionnaire’. Whilst they were completing this
questionnaire the experimenter left the room and a ‘thief’, played by one of the actors, entered the room and attempted to steal a laptop computer. Afterwards, participants were informed of the true purpose of the study and gave a free recall description of the thief. These descriptions were used to create description matched lineups, described in more detail in the Materials section above.

**Session 2 procedure**

Participants attended Session 2 individually to carry out an identification procedure. Instructions emphasised that the culprit ‘may or may not be present’ in the lineup, and that if the witness could not recognise the culprit, they must state that they could not make an identification.

Participants watched the entire VIPER parade through twice from start to finish. They were then asked if they wished to see any of the lineup members again before making their decision, and were allowed to see any number of items again, as many times as they requested. Finally their decision was recorded. This procedure follows the current PACE code of practice.

**Independent ratings of experimental stimuli**

Five independent raters (all female, mean age = 27) judged the degree of match between the witness’ descriptions of the culprit and the lineup members. Raters were asked to judge how well the written description described the person using a 7-point Likert scale. Each description was shown once paired with each foil from the lineup used for that particular witness as well as with the culprit and the innocent suspect. In all, each rater made 2000 judgements. To avoid fatigue effects, ratings were made on five separate occasions, with order of comparisons randomised across participants.

Nine raters (eight female, mean age = 25) judged similarity between suspects and lineup members. Pairs of still VIPER images were presented and raters were asked to judge the ‘degree of similarity’ between the two people on a 7-point Likert scale. The pairs always included one suspect (either an innocent suspect or a culprit), and a foil used in one of the lineups for that suspect. All of the possible pairs of suspects and relevant foils were presented for rating, a total of 1632 pairs (note that the CD lineups, which were individually chosen for each participant, therefore substantially outnumbered the SR lineups). To avoid fatigue effects, ratings were made on two separate occasions with order of comparisons randomised across participants.

**RESULTS**

**Culprit present lineups**

Figure 1 shows the percentage of the total number of participants who saw a lineup of each type and made a correct identification of the culprit, an incorrect rejection of the lineup or a mistaken identification of a foil.

The numbers of participants making each response were analysed by use of a hierarchical loglinear analysis. A saturated model was constructed in which the outcome of
each lineup, foil selection method and image format were entered as variables. Components of this model were removed using backward elimination. Probability for removal in the model was $p < 0.05$. No significant effects were present. Subsequent $\chi^2$ tests confirmed this pattern for the main effects, with no significant difference in the distributions of responses between SR and CD lineups ($\chi^2 (1) = 0.19, p > 0.1, \Phi = 0.07$), or between the still and moving lineups ($\chi^2 (1) = 0.62, p > 0.1, \Phi = 0.08$). Post-hoc power analysis indicated that this sample had a power of 0.85 to detect a medium effect size ($\Phi = 0.3$).

Culprit absent lineups

Figure 2 shows the outcomes for participants viewing CA lineups. The data were analysed using a hierarchical loglinear analysis, conducted with the same parameters described for CP lineups. No significant effects were present. Subsequent $\chi^2$ tests confirmed this pattern, with no significant difference in the distributions of responses between SR and CD lineups ($\chi^2 (1) = 0.56, p > 0.1, \Phi = 0.16$) or between still and moving lineups ($\chi^2 (1) = 1.15, p > 0.1, \Phi = 0.11$). Post-hoc power analysis indicated that this sample had a power of 0.85 to detect a medium effect size ($\Phi = 0.3$).

Independent ratings of match of lineup foils to witness descriptions

The mean of the five independent raters’ ratings of the match of lineup items with witness descriptions of the culprit for SR lineups was 3.17 (SD = 1.23), whilst for CD lineups the

\(^2\)Expected frequencies in the ID foil category were below five. Therefore loglinear and $\chi^2$ analyses were conducted with the two separate error categories collapsed into one general error category.

\(^3\)Expected frequencies in the ID innocent suspect category were below five. Therefore loglinear and $\chi^2$ analyses were conducted with the two separate error categories collapsed into one general error category.
equivalent mean was 3.55 (SD = 1.38). This difference was statistically significant ($t(4) = 3.9, p < 0.05, d = 0.28$). All of the independent raters rated CD lineup members as being better matches to the witness descriptions than SR lineup members. Averaged inter-rater reliability across all ratings (intraclass correlation coefficient) was 0.70.

**Independent ratings of foil-suspect similarity**

Means of nine independent raters’ ratings of foil-suspect similarity are presented in Table 1. Raters rated SR lineup foils to be significantly more similar to the suspect than CD lineup foils across all lineups ($t(8) = 8.60, p < 0.001, d$ (Cohen’s effect size for $t$-tests (Cohen, 1977) = 0.97)), for CP lineups ($t(4) = 3.81, p < 0.001, d = 0.53$) and for CA lineups ($t(4) = 8.71, p < 0.001, d = 1.38$). The SR foils were rated as being significantly more similar to the suspects (whether guilty or innocent) than were the CD foils (mean for SR foils = 3.31 (SD = 0.68), mean for CD foils = 2.98 (SD = 0.63)). Averaged inter-rater reliability across all ratings (intraclass correlation coefficient) was 0.88.

**DISCUSSION**

There were no significant differences between SR and CD parades, either in the CP or CA conditions. Effect sizes related to the foil selection manipulation were small: to recruit
enough participants for an effect of these magnitudes to achieve observed power of 0.80 would require a sample of 1602 participants for CP lineups and 307 for CA lineups. Therefore, we conclude that any differences in number of correct identifications in CP lineups and number of innocent suspect identifications in CA lineups, if present at all, were too small to detect with the current experimental design.

This experiment was designed to probe for main effects: the number of observations was sufficient to indicate the presence of medium effects with a power (β) of over 0.80. The critical \( \chi^2 \) value in both CP and CA analyses was 3.84. A very similar methodology was used by Valentine et al. (in press), who observed significant effects of viewing procedure in CP lineups (sequential vs. simultaneous: \( \chi^2 = 9.82 \)) and of image movement in CA lineups (\( \chi^2 = 5.41 \)): the current experiment had the power to detect effects of comparable size. However, the observed effect sizes were small, therefore the lack of a systematic effect of description matching or of moving images is not conclusive. Even a small effect, though, may have a substantial impact when amplified across the large number of lineup procedures carried out across various international jurisdictions, and so is potentially non-trivial. Further larger scale studies may confirm or reject the presence of such effects. We would argue, however, that in this case the most likely explanation of the lack of significant differences between CD and SR lineups was that there was minimal or no benefit of CD lineups.

Independent ratings indicated that the foils in CD lineups matched witness descriptions of the culprit better than did foils in SR lineups. Furthermore, resemblance to the suspect was higher in the SR than in the CD conditions. The fact that CD foils were rated as less similar to suspects than SR foils confirms that CD lineups were more heterogeneous than the SR lineups. Explanations of our null result that rely on claiming that the foils were poor matches to descriptions or to suspect appearance can thus be ruled out. Despite this evidence that manipulations were implemented successfully, we observed no significant differences in outcome according to foil selection strategy.

Ratings of similarity in the SR condition were not particularly high, leaving open the possibility that the failure to find a reduction of hits for SR lineups compared to CD lineups occurred because our SR stimuli were not similar enough in appearance. The extent that this is true is difficult to evaluate, because the relationship between ‘objective’ similarity and subjective assessments of similarity based on a fairly abstract scale is difficult to estimate. Whilst acknowledging this possibility, however, it is also important to note that all nine raters rated SR lineups to be more similar than CD lineups (i.e. there was no overlap in the mean ratings), both for CP and CA lineups, and these comparisons were highly significant. All five raters of the descriptions rated the CD descriptions as being better matches to foils in CD lineups as to foils in SR lineups, and again, this difference was significant. Inter-rater reliability measures were high for both sets of ratings. Furthermore, the SR lineups were produced following the current procedure in the UK, using the police database. Therefore, the SR lineups tested here reflected the level of resemblance found in an operational context.

Luus and Wells (1991) and Wells et al. (1998) identify inaccurate descriptions and over-detailed descriptions as being potential sources of bias in CD lineups. To exclude the possibility that these types of descriptions were exerting undue influence on our data, we removed participants from the analysis in the event that raters rated the mean foil—description agreement as being extremely low (i.e. in the bottom 5% of scores) and also participants whose descriptions contained more than eight useful items, not including default values of sex and ethnicity (i.e. the most detailed 5%). The pattern of results remained unchanged.
To our knowledge, this is the second report to find minimal effects of foil selection strategy: Tunnicliff and Clark (2000) report two experiments in which they found no effects. However, our data are discordant with the results of Juslin et al. (1996), Lindsay et al. (1994) and Wells et al. (1993). In explanation of their results, Tunnicliff and Clark (2000) suggest that the difference between their data and that of Wells et al. and Lindsay et al. might be related to the delay between presentation and recognition. In the case of the earlier studies, this was in the order of minutes, whilst in their studies (and also in our current experiment) delays were in the order of several days. Juslin et al. manipulated delay (1 hour vs. 1 week), but they did not specifically report the results of the retention interval manipulation with regard to foil selection strategy.

Delay duration was not the only difference between the current study and those in which CD–SR differences have been observed. As described in the introductory section, VIPER lineups differ from photospread lineups by virtue of their sequential nature, use of video and use of a large but constrained database of distractors. Any one of these factors might account for the difference between our data and that of other authors supporting the use of description matching. The requirement to use the VIPER database may have been particularly critical, because foils used by Wells et al. (1993) had a high degree of similarity in contrast to the heterogeneity of their description matched foils (see Clark, 2003 for a discussion): it may not be possible to replicate such a degree of similarity within an operational database such as that used by VIPER. However, it is clearly possible to produce match-to-suspect lineups that are qualitatively distinct from match-to-description lineups, as demonstrated by the independent ratings in the current study. Therefore we echo the conclusions of Tunnicliff and Clark (2000) that there is nothing inherent in the suspect—matched foil selection strategy that produces low rates of correct identification by necessity.

Three separate experiments (two by Tunnicliff and Clark, 2000, and the current study) have failed to observe statistically significant benefits for description matching. Particularly noteworthy is the fact that these three studies are by some margin the most ecologically valid, both employing staged crime scenarios and realistic delay durations. Our study extends this by also using a police identification system in exactly the manner as would be employed operationally. Therefore, it seems likely that the benefits reported for description matching in laboratory experiments may not transpose to the operational context.

Valentine et al. (in press) observed a small effect whereby there were more incorrect identifications on CA lineups when participants saw still full—face video images compared to moving images. This effect was not reliably replicated in the current study, although the pattern of data did show a trend in the same direction. When added to data from the Valentine et al. (in press) experiment from the identical SR experimental condition, the comparison remained significant ($\chi^2(1) = 5.48, p < 0.05, \Phi = 0.17$). Hence, although there was no benefit observed for using video images in the current study on its own, combining the data with that from the Valentine et al. study suggests that a small benefit was observable.

**Implications for policy**

Wells et al. (1998) recommended that CD matching be used in place of SR matching in US practice, and this was incorporated into a US Justice Department guide of best practice (Technical Working Group on Eyewitness Evidence, 1999). However, taken alongside
Tunnicliff and Clark’s (2000) data, our data suggest that such conclusions may be premature. Despite the fact that suspect matching is associated with fewer culprit identifications in highly controlled laboratory experiments, it seems increasingly to be the case that such effects either do not generalise to more ecologically valid experiments, or are too small to be observed in such experiments. By inference it would seem unlikely that such effects would generalise to operational use.

Given that the clear benefits claimed for description matching now seem questionable, then its disadvantages become more relevant. These include the situations where description matching is problematic: when descriptions are inaccurate, when they are over-detailed or when they are poorly specified. Furthermore, there remains the complex issue of how to deal with incomplete descriptions and default values (Lindsay et al., 1994), and potential proneness to the backfire effect (Clark, 2003).

PACE dictates that selection of foils for lineups should be based on SR. The current data provide no basis to change this. Furthermore, the fact that VIPER uses a central database for foil selection may impose an upper limit on the degree of similarity attainable in SR lineups, and hence mitigate against the logically extrapolated ‘lineup of clones’ argument.

PACE recommends that a moving image be preferred to a still one where practicable. Our current data showed no reliable evidence of differences in lineup performance as a result of using still or moving images in VIPER lineups. In our previous study (Valentine et al., in press), we found a fairly small but nonetheless significant improvement in lineup rejection for moving compared to still CA lineups, and this effect is still present when the data from both studies are combined. Therefore, although our data across both studies favour the use of moving image lineups if possible, we see no strong argument to make the preference in the code of practice for moving images mandatory in all cases without exception. This recommendation may also be relevant to readers with an interest in North American jurisdictions, where the use of static images is typical.

ACKNOWLEDGEMENTS

We would like to thank the Nuffield Foundation for funding this research project (Grant no. OPD/00227/G), and West Yorkshire Police and the staff at the VIPER bureau who compiled the lineups. We would also like to thank Kate Houston for her valuable assistance in collecting data for this study.

REFERENCES


Selecting lineup foils


