Law and Human Behavior, Vol. 24, No. 2, 2000

Selecting Foils for Identification Lineups: Matching Suspects or Descriptions?

Jennifer L. Tunnicliff[1] **and Steven E. Clark**[1]

Two experiments directly compare two methods of selecting foils for identification lineups. The suspect-matched *method selects foils based on their match to the suspect, whereas the* description-matched *method selects foils based on their match to the witness's description of the perpetrator. Theoretical analyses and previous results predict an advantage for description-matched lineups both in terms of correctly identifying the perpetrator and minimizing false identification of innocent suspects. The advantage for description-matched lineups should be particularly pronounced if the foils selected in suspect-matched lineups are too similar to the suspect. In Experiment 1, the lineups were created by trained police officers, and in Experiment 2, the lineups were constructed by undergraduate college students. The results of both experiments showed higher suspect-to-foil similarity for suspect-matched lineups than for description-matched lineups. However, neither experiment showed a difference in correct or false identification rates. Both experiments did, however, show that there may be an advantage for suspect-matched lineups in terms of no-pick and rejection responses. From these results, the endorsement of one method over the other seems premature.*

INTRODUCTION

The identification of a suspected criminal by an eyewitness provides extremely compelling evidence for conviction in criminal cases (Huff, Rattner, & Sagarin, 1986; Greene, 1988; Loftus, 1974; Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998). Thus, it is imperative that this evidence be extremely reliable.

A standard identification procedure uses a lineup, which typically consists of one suspect and some number of known nonsuspects, often referred to as *foils* or *distractors.* The foils play a critical role in the function of lineup identification procedures. One purpose of the foils is to make it unlikely that the witness will identify the suspect through simple or sophisticated guessing strategies (see Luus & Wells, 1991, for an extended discussion regarding the role of foils in lineups). The

[1] Department of Psychology, University of California, Riverside, California.

0147-7307/00/0400-0231$18.00/1 © 2000 American Psychology-Law Society/Division 41 of the American Psychology Association

main question addressed in the present research is how to select the foils for the lineup.

One strategy for selecting foils is to select foils who look like the suspect. A national survey by Wogalter, Malpass, and Berger (1993) reported that 83% of police officers construct lineups using this *suspect-matched* method. The suspect-matched method is appealing for two reasons: (1) It provides a foundation for the prosecution to argue that the witness who has identified the suspect can distinguish between the perpetrator (assuming the suspect is guilty) and several people who look very similar to the perpetrator, and (2) it should protect innocent suspects from false identification by surrounding them with look-alikes.

As straightforward as the match-to-suspect rule may appear, some researchers have argued that it may have potentially serious problems, and they have proposed an alternative method for selecting lineup foils—that the foils should be matched not to the photograph of the suspect, but rather to the description of the perpetrator given by the witness (Brooks, 1983; Luus & Wells, 1991; Navon, 1992; Wells et al., 1998). This alternative strategy is referred to as a *description-matched* method of foil selection. In this article, we review the potential problems with suspect-matched lineups and the arguments in favor of the description-matched method of selecting foils.

Potential Problems with the Suspect-Matched Method

The match-to-suspect method has been criticized on three grounds: (1) It does not give a clear definition of similarity, (2) it may reduce the likelihood of correctly identifying the suspect when he or she is guilty, and (3) it may produce a high rate of false identifications when the suspect is innocent. If these criticisms are valid, those police officers who use the match-to-suspect method, which may be as many as 83% of the police officers in the United States, may be using a flawed procedure. Each of these criticisms is discussed in greater detail below.

No Clear Definition of Similarity

If the task is to pick foils that look similar to the suspect, we must answer the question, *How similar is similar enough?* Luus and Wells (1991) argue that there is no clear answer to the ‘‘similar enough’’ question. In the simplest case in which two individuals are described in terms of features (such as hair color, height, weight, etc.), how many of those features must match before one can assert that Individual A and Individual B are similar enough? For any two individuals who match on some number of features, we can surely find some feature on which they do not match—unless the two are clones, in which case the problem has been forced into absurdity. Somewhere between an unfair lineup and a lineup of clones there is presumably some appropriate level of similarity; however, the match-to-suspect method provides no standards for determining that level.

Decrease in Correct Identification Rate

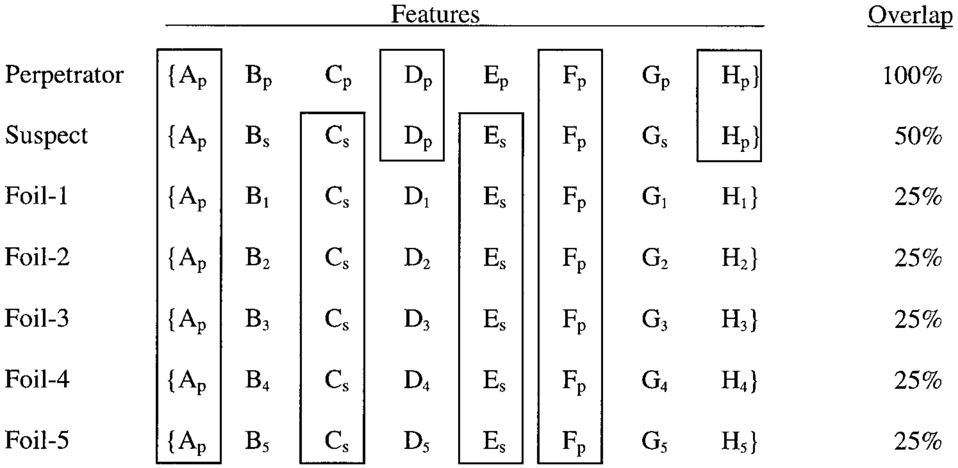
Luus and Wells (1991) argue that the lack of a clear standard for similarity may lead to decreased rates of correct identification. Specifically, if lineup fairness is defined in terms of the similarity of the foils to the suspect, a conscientious police officer, in attempting to create a very good lineup, may select foils that are so similar to the suspect that even a witness with a reasonably good memory of the perpetrator might have difficulty distinguishing the suspect from the foils. Such lineups would serve to protect not only innocent suspects, but guilty ones as well.

High Rate of False Identification

Navon (1992) derived predictions from a mathematical model and showed that selecting foils based on their similarity to the suspect might backfire. Specifically, in the case in which the suspect is innocent, the suspect-matched method of selecting foils may produce a lineup in which the innocent suspect is more similar to the real perpetrator than any of the foils. The assumptions of this model are quite simple, and if correct, they show that the innocent suspect is the *most* likely person to be picked from a lineup in which the foils are selected based on their similarity to the suspect.

Consider the case in which an innocent person is suspected in part because he matches the description of the perpetrator given by the witness. He is then placed in a lineup along with five foils who are selected based on their match to the suspect. The important thing to note here is that the suspect and the foils are included in the lineup based on different criteria. The suspect is in the lineup because he is similar to the perpetrator, but the foils are in the lineup because they are similar to a person who is *not* the perpetrator, but only similar to the perpetrator. Thus, the innocent suspect is one level removed in terms of similarity to the perpetrator (because he fits the description of the perpetrator), but the foils are two levels removed (because they match the person who matches the description).

The problem is illustrated in Fig. 1, a simplified version of Navon's derivations. The letters A, B, C, . . . , H denote a set of dimensions that may used in determining



	Features								Overlap
Perpetrator	$\{A_p$	B_p	C_p	D_p	E_p	F_p	G_p	$H_p\}$	100%
Suspect	$\{A_p$	B_s	C_s	D_p	E_s	F_p	G_s	$H_p\}$	50%
Foil-1	$\{A_p$	B_1	C_s	D_1	E_s	F_p	G_1	$H_1\}$	25%
Foil-2	$\{A_p$	B_2	C_s	D_2	E_s	F_p	G_2	$H_2\}$	25%
Foil-3	$\{A_p$	B_3	C_s	D_3	E_s	F_p	G_3	$H_3\}$	25%
Foil-4	$\{A_p$	B_4	C_s	D_4	E_s	F_p	G_4	$H_4\}$	25%
Foil-5	$\{A_p$	B_5	C_s	D_5	E_s	F_p	G_5	$H_5\}$	25%

Fig. 1. Illustration of Navon's predictions. Features are denoted by letters A–H. Subscript p, perpetrator; subscript s, suspect. Vertical boxes denote the overlap of features across perpetrator, suspect, and foils.

similarity; these might include features such as hair color, eye color, height, ethnicity, etc. The subscripts p and s denote the feature values on those dimensions for the perpetrator and the innocent suspect, respectively. We make the assumption that the amount of feature overlap between the perpetrator and the suspect is the same as that between the suspect and any foil. For this example, it is assumed that the overlap is on half of the features, but any overlap value produces the same outcome. Thus, the suspect matches the perpetrator on half of the features (denoted A, D, F, and H), and each foil matches the suspect on half of the features (features A, C, E, and F), but each foil matches the perpetrator on one-fourth of the features (A and F). Navon's analysis has been shown to hold up in the predictions of a computer simulation model of eyewitness identification developed by Clark (1999), and may generalize to an entire class of similarity-based models of memory (Clark & Gronlund, 1996).

Description-Matched Lineups: A Possible Solution

The recommended solution to these potential problems is to select foils based on their match to the description of the perpetrator, rather than based on their similarity to the photograph of the suspect. The difference between suspect-matched and description-matched lineups is the target stimulus to which foils are compared (suspect photo or witness description). Why should this small procedural difference solve the problems previously noted?

First, description-matched lineups provide an answer to the ‘‘similar-enough’’ question; the matching stops at the end of the witness description. Second, because witness descriptions are typically brief, it is difficult to overmatch the foils to a level of similarity that makes the lineup alternatives indistinguishable. Third, in the case in which an innocent person becomes a suspect due to his or her matching the witness's description, both the innocent suspect and the foils are in the lineup based on the same criterion. Thus, the innocent suspect should be no more likely to be picked than any of the foils. This equality between the suspect and foils may be the most important foundational principle favoring description-matched lineups over suspect-matched lineups.

Relevant Experiments

The criticisms suggest that the suspect-matched method of selecting foils is flawed in such a way that it produces unnecessarily low correct identification rates and high false identification rates. The question is, do these problems actually occur? This question was addressed by two experiments, one by Wells, Rydell, and Seelau (1993) and another by Lindsay, Martin, and Webber (1994).

Wells et al. (1993) used a procedure in which witnesses saw a staged crime—a person stealing money from a university laboratory. Each witness was shown either a suspect-matched lineup or a description-matched lineup. The results showed that correct identification rates were higher for description-matched than for suspect-matched lineups, consistent with predictions by Luus and Wells (1991). However, there was no difference for false identifications. In the suspect-matched lineups, the false identification rate for the innocent suspect was not higher than the average identification rate of the other foils. The equality of the false identification rates, although inconsistent with Navon's (1992) analysis, suggests a strong advantage for description-matched lineups: Compared to suspect-matched lineups, description-matched lineups allowed for a substantial increase in correct identification rates with no increase in false identification rates.

Navon's (1992) analysis can also be evaluated by looking at the conditional probability of picking the innocent suspect, given that the witness makes any pick. If the innocent suspect looks more like the perpetrator than anyone else in the lineup, the conditional probability should be greater than .167 (for a six-person lineup). Wells et al. (1993) showed conditional probabilities of .20 and .278 for suspect-matched and description-matched lineups, respectively. Although both probabilities were greater than chance (.167), neither deviation was statistically reliable ($z = .447$ and $z = 1.265$, respectively).

A similar pattern of results was also shown by Lindsay et al. (1994), although the correct identification advantage for the description-matched lineups over the suspect-matched lineups was not as large as was shown by Wells et al. (1993). However, Lindsay et al. showed a surprising pattern of conditional probabilities. The conditional probabilities of picking the innocent suspect were .176 and .563 for suspect-matched and description-matched lineups, respectively. These results are in the opposite direction of the prediction, and the conditional probability in the description-matched lineups did differ significantly from chance ($z = 4.249$, $p < .001$).

The results of these experiments suggest that the selection of foils based on their match to the suspect's photograph can result in a lineup in which the foils are unnecessarily similar to the suspect, and moreover, that this ''gratuitous'' similarity can result in lower rates of correct identification relative to description-matched lineups. These results were the motivation for a recent recommendation by Wells et al. (1998) that ''distractors should not necessarily be selected so as to look like the suspect, but instead should be selected so that they fit the description. . . . Selecting distractors so as to resemble the suspect is not a desirable practice'' (p. 632). Wells et al. (1998) further suggest that the suspect-matched method could potentially meet certain standards (i.e., mock-witness test), although ''such practices might create undue homogeneity and interfere with the recognition of the actual culprit'' (p. 632). However, demonstrations from previous studies that it *can* happen do not imply that it *must* happen by *necessity.*

Assuming that the overmatching of foils to the suspect is not a result that must occur by necessity, and that there may be some variability in the results (which there appears to be, comparing results from Lindsay et al., 1994, and Wells et al., 1993), an important question is: Does it happen in the lineups that are constructed by police officers? Do police officers select foils that are unnecessarily similar to the suspect? Also, could police officers create better lineups (higher correct identifications and lower false identifications) by selecting foils based on their similarity to the description of the perpetrator rather than their similarity to the suspect's photograph?

To address these questions, in Experiment 1, the lineups were created by police officers who had at least 3 years experience in making lineups as a part of their normal duties. To implement the two different foil selection methods, each officer was given either a photograph of the suspect or a witness's description of the perpetrator (but not both). The officers constructed the lineups using foils from their own booking photo database. With these procedures Experiment 1 addresses two questions: (1) Will the foils in suspect-matched lineups be more similar to the suspect than the foils in description-matched lineups? (2) Will there be differences in the patterns of identification responses for these two methods of foil selection? The second question focuses on, but is not limited to, correct and false identification rates.

EXPERIMENT 1

Experiment 1 used a staged crime procedure, followed by the presentation of a lineup 2 weeks later. The lineups were created by police officers who selected foils based on their match to the suspect photograph or to the witness description.

As in previous experiments, each lineup contained either a photograph of the thief or a photograph of an innocent suspect (who was determined by police officers to fit the description of the thief given by the witness). These conditions are termed *perp-present* and *perp-absent,* respectively. Both perp-present and perp-absent conditions are necessary to determine whether increases in correct identification rates are tied to increases in false identification rates.

Each witness gave an identification response and two measures of confidence, one just after the staged crime and one just after the identification response. A second group of participants (who did not see the crime) rated the lineups in terms of the similarity of the foils to the suspect.

Method

Participants

A total of 182 people participated, of whom 128 were witnesses to the staged crime and an additional 54 provided similarity ratings for the lineups. Participation was arranged either through an introductory psychology course requirement or through campus advertisements.

Materials

All photographs in the lineups were police mugshots from the San Bernardino County Sheriff's Department. The confederate-perpetrators had mugshot booking photographs taken at this facility for use in the photo lineups. Two different perpetrators appeared throughout this experiment with equal frequency. One perpetrator was a 26-year-old White male, 5′ 7″ tall, weight 160 lb, with short brown hair and hazel eyes. The other perpetrator was a 24-year-old White male, 6′ 1″ tall, weight 190 lb, with long (below the shoulders) sandy blond hair and blue eyes. Each lineup consisted of a photograph of the suspect and five foil photographs that were attached to lineup mounts.

Procedure

Session 1

An average of seven participants per group was seated at a conference table in a 9 foot × 20 foot experiment room. Participants completed the Texas Social Behavior Inventory (Helmreich & Stapp, 1974),[2] after which the experimenter left the room stating, ‘‘I will be right back.’’ When the experimenter had been absent for approximately 1 min, the perpetrator entered the room, circled around the participants, making eye contact, and asked, ‘‘Where did the person go who is running this experiment?’’ The perpetrator then searched for and removed an envelope marked ‘‘participant money’’ from a bookshelf, said ‘‘See ya later!’’ and then left the room. The event lasted approximately 30 sec.

Having been told of the theft on her return, the experimenter pretended to phone the campus police. The experimenter then announced to the participants that the campus police were on their way to question them about the missing money and obtain their descriptions of the thief. The experimenter explained to participants that the police would like for them to write down individual descriptions of the thief while they were waiting. After all participants had written descriptions, the experimenter told the participants that the theft was staged.[3] The experimenter then prompted participants to write down more elaborate descriptions of the thief on a standard description form.

Participants were asked to rate their confidence in their ability to later identify the perpetrator using a scale of 1 (not at all confident) to 5 (most confident). Participants reported other information such as seat locations, whether they got a ‘‘good’’ look at the perpetrator, and whether they had talked to the perpetrator during the interaction. Participants were then dismissed, with no instructions that would indicate that they would view a lineup when they returned. All participants in this session were run within a 3-day time period in order to prevent discussion of the experiment with other potential participants.

Photo-Lineup Construction. Descriptions and suspect photographs were given to the Sheriff’s Department, where 96 detectives participated in creating lineups.[4] These detectives had at least 3 years of experience constructing lineups. For the current study, each detective made only two lineups, one for each perpetrator. This was done to avoid contamination across witnesses or experimental conditions. For example, a detective making several lineups all based on slightly different descriptions of the same suspect might create lineups based on an emerging prototype of the perpetrator. Such a situation could also result in a witness’s being presented with a lineup in which the foils were selected in part from another witness’s description.

Each detective had a different task because there were four different conditions

[2] The Texas Social Behavior Inventory, developed by Helmreich and Stapp (1974), is a 16-item questionnaire used to measure self-esteem.

[3] At this point if we had not partially debriefed participants, it would have become obvious that the theft was staged when we handed participants the perpetrator description forms.

[4] Sixteen detectives were required for each of the four conditions. Thirty-two additional detectives were needed in order to select innocent suspects in the perp-absent conditions.

and 128 different descriptions (some tasks were only slightly different). In suspect-matched conditions, detectives selected five foils based on the similarity to the suspect photograph. Detectives never saw the witness description in this condition. In description-matched conditions, detectives selected five foils based on the match to the witness description. In this condition, detectives did not see the photograph of the suspect. These procedures made it impossible for a detective in one condition to create a lineup based on a different condition.

For perpetrator-absent conditions, the innocent suspect was selected by having one detective select a suspect photograph based on the witness description and then having another detective use this photograph or the witness description (depending on the condition) to select the remaining five lineup foils. The detectives chose the position of the suspect and foils in each lineup.[5]

Session 2

Each participant returned 2 weeks after the first session to view a lineup. Four different experimenters administered the lineups. In order to avoid experimenter bias effects (Rosenthal, 1966), the experimenters never saw the perpetrators or the lineups themselves. Before viewing the lineup, the participant was read a standard lineup admonition that stated:

> In a moment I am going to show you a group of photographs. This group of photographs may or may not contain a picture of the person who committed the crime now being investigated. Keep in mind that hair styles, beards, and moustaches may be easily changed. Also, photographs may not always depict the true complexion of a person—it may be lighter or darker than shown in the photo. Pay no attention to any markings or numbers that may appear on the photos or any other differences in the type or style of the photographs. When you have looked at all the photos, tell me whether or not you see the person who committed the crime. Do not tell other witnesses that you have or have not identified anyone.

Participants then had the opportunity to select a photograph from the lineup presented to them. For those participants who did not make a selection, their responses were recorded as no-pick. These participants were then asked the reason for their no-pick response: Was it because they *did not know* or because they believed that the *perpetrator was not present* in the lineup? If participants made more than one selection from the lineup initially, both responses were recorded. In this case, they were later asked to narrow their response down to one photograph. Because only six participants made multiple picks, we did not separately analyze these results.[6] Nonchoosers were later required to select the best match from the lineup; however, these results were analyzed separately. Finally, all participants were asked to rate their confidence in their final answer on a scale of 1 (not at all confident) to 5 (most confident). Participants were then debriefed and dismissed.

Similarity Ratings. The lineups were evaluated by a separate group of participants who were selected from the same pool as the witnesses, but who had not

[5] An examination of the suspect position in the lineup showed that no position was chosen by detectives more often than any other, $\chi^2(1, N = 128) = 1.219$, ns). The frequencies were 11, 13, 11, 10, 14, and 14, for positions 1-6 respectively.

[6] However, we did note that three were from description-matched lineups and three were from suspect-matched lineups.

been witnesses to the staged crime. These participants rated the similarity between the suspect and each foil in each lineup on a 1 (least similar) to 7 (most similar) scale. Because the number of lineups was fairly large, each subject gave ratings for either the perp-present or the perp-absent lineups ($N = 27$ in each group). Thus, each rater provided 320 ratings (64 lineups with five foils per lineup).

Results and Discussion

Results are presented for several sets of analyses for (1) similarity ratings, (2) identification responses, (3) witness descriptions, and (4) confidence data.

Similarity Ratings

Each lineup provided five similarity ratings, comparing each foil to the suspect. These five ratings were then averaged within each lineup, and then averaged across lineups separately for the suspect-matched and description-matched lineups. Mean ratings for the perp-present lineups were 2.44 ($SD = .938$) for suspect-matched and 2.11 ($SD = .811$) for description-matched. Perp-absent mean ratings were 3.13 ($SD = .928$) and 2.82 ($SD = .881$). These analyses showed higher similarity ratings in suspect-matched than description-matched lineups for both perp-present, $t(26) = 3.327$, $p < .005$, and perp-absent, $t(26) = 5.046$, $p < .001$, lineups.[7] A moderate effect size of .546 and large effect size of .703, respectively, were obtained. Thus, the similarity ratings are consistent with Luus and Wells' (1991) prediction. Will these similarity differences produce differences in the identification responses given by witnesses?

Identification Responses

The witness responses were classified as one of the following: identification of the suspect, identification of a foil, or a no-pick. The no-pick responses were subdivided into ‘‘don't know'' responses and rejections of the lineup (perpetrator not in the lineup). These response probabilities are shown in Table 1 for suspect-matched and description-matched lineups and for perp-present and perp-absent conditions.

Chi-square tests were conducted to evaluate differences between suspect-matched and description-matched lineups. Prior to the main analyses, preliminary analyses showed that the overall patterns of responses were different for perp-present and perp-absent lineups, $\chi^2(6, N = 128) = 33.15$, $p < .001$; thus, subsequent analyses were conducted separately for perp-present and perp-absent conditions. There was no difference between perpetrators, perp-present $\chi^2(2, N = 64) = 3.429$, ns, perp-absent $\chi^2(2, N = 64) = 2.669$, ns, so subsequent analyses will collapse over the two perpetrators.

Two chi-square tests were conducted for perp-present and perp-absent conditions to determine if the overall patterns of responding were different for suspect-matched and description-matched lineups. For these tests, the no-pick responses

[7] To ensure our results were not due simply to the difference between parametric and nonparametric statistical tests, the Wilcoxon Signed Ranks Test for rank-order data was also calculated. This test yielded significant differences for perp-present, $z\ (26) = -3.857$, $p < .001$, and perp-absent $z\ (26) = -3.709$, $p < .001$) conditions.

Table 1. Identification Responses for Experiment 1

	Perpetrator present		Perpetrator absent	
	Suspect matched (%)	Description matched (%)	Suspect matched (%)	Description matched (%)
Suspect	53.1	53.1	3.1	12.5
Foil	25.0	15.6	31.3	34.3
No-pick	21.8	31.3	65.5	53.1
Don't know	9.3	3.1	9.3	18.7
Perpetrator not there	12.5	28.1	56.2	34.3

Note: A comparison of suspect-matched and description-matched conditions for suspect identifications, foil picks, and no picks. Also included is a breakdown of the no-pick responses: ''don't know'' and perpetrator not in the lineup.

were not subdivided because to do so would have produced several cells with expected frequencies less than 5. These tests indicated that the overall patterns of responses were not different for suspect-matched and description-matched lineups, either for perp-present, $\chi^2(2, N = 64) = 1.22$, ns, or perp-absent, $\chi^2(2, N = 64) = 2.27$, ns, lineups.

We compared suspect-matched and description-matched lineups for each of the four response categories (subdividing no-picks into ''don't know'' and ''reject'' responses) for perp-present and perp-absent lineups. None of these comparisons reached statistical significance, although a reliable interaction occurred for rejection responses.

Correct and False Identification. Table 1 shows that the proportion of correct identifications was identical (.531) for suspect-matched and description-matched lineups, $\chi^2(1, N = 64) = 0$, ns. False identification rates were also similar for suspect-matched and description-matched lineups, $\chi^2(1, N = 64) = 1.952$, ns, although this analysis should be interpreted cautiously because the numbers of cases in these two cells were both very small. Only one subject falsely identified the innocent suspect in suspect-matched lineups, and only four subjects made false identifications in the description-matched lineups. This pattern of results is in the opposite direction of the predictions by Navon (1992).

We also calculated for the perp-absent lineups the conditional probability of picking the suspect given that the witness picked anyone. In a fair six-person lineup, these probabilities should be .167. These probabilities were .091 for suspect-matched lineups and .267 in description-matched lineups, neither of which deviated significantly from chance ($z = .447$ and $z = 1.265$, ns).

In light of the similarity ratings that show higher foil-to-suspect similarity in suspect-matched than in description-matched lineups, the absence of any differences in the identification results is somewhat surprising. We return to this dissociation of similarity and identification data later.

Foil Identifications. For perp-present lineups, foils were identified slightly more often in suspect-matched lineups than in description-matched lineups, consistent with the similarity-rating data. However, this small difference was not reliable,

$\chi^2 = .739$, ns. In contrast, for perp-absent lineups, the foil identification rates were virtually even in suspect-matched and description-matched lineups.

No-Pick and Lineup Rejections. For description-matched lineups, the proportions of no-pick responses were not significantly different for perp-present and perp-absent lineups, $\chi^2 = 3.139$, ns. However, for suspect-matched lineups, no-pick responses were much more likely for perp-absent lineups than perp-present lineups, $\chi^2 = 12.444$, $p < .001$. The test of the interaction (Langer & Abelson, 1972), however, was not significant ($z = 1.428$).

This pattern was particularly strong for lineup rejections. For description-matched lineups, witnesses were about as likely to reject the lineup for perp-present lineups as for perp-absent lineups, $\chi^2 = .291$, ns. However, for suspect-matched lineups, rejections occurred significantly more often for perp-absent lineups than for perp-present lineups, $\chi^2 = 13.576$, $p < .001$. The interaction, tested in the same way, was reliable ($z = 2.298$, $p < .05$).

Including Forced-Choice Data. We reanalyzed the results after adding the data from subjects who were forced to make a pick. The correct identification rates increased for suspect-matched (.688) and description-matched (.625) lineups, but again the difference was not statistically reliable, χ^2 (1, $N = 64$) = .276, ns. Likewise for perp-absent lineups, the rates for false identifications increased when subjects were forced to pick, but the proportions did not differ for suspect-matched (.188) and description-matched (.250) lineups, $\chi^2(1, N = 64) = .360$, ns. Two aspects of these results should be noted. First, the 18.8% false identification rate is not different than the 16.7% rate that would be expected by chance guessing, and again, the trend in the results is in the opposite direction: false identification rates were lower (though not reliably so) for suspect-matched lineups than for description-matched lineups.

Functional Size. Functional size was calculated using the same method as in Wells et al. (1993), in which functional size is given by the ratio of the number of people who pick anyone divided by the number of people who pick the suspect in perp-absent lineups. A fair and unbiased lineup should have a functional size of six. If there is a bias in suspect-matched lineups, as predicted by Navon (1992), and if that bias is reduced in description-matched lineups, then functional size for suspect-matched lineups should be less than six, and functional size should be larger for description-matched lineups. The present results are inconsistent with that analysis, with functional sizes of 11 and 3.75 for suspect-matched and description-matched lineups, respectively. Again, however, the numbers of observations are quite small, so these functional sizes should be viewed cautiously.

Witness Descriptions

Initial descriptions obtained through unprompted free recall contained an average of 6.27 descriptors. When participants were later prompted with a description sheet, their descriptions increased to 11.21 items. Many of the items described details that would not be relevant for foil selection (details about clothing, jewelry, etc.); with such items excluded, the descriptions contained an average of 7.94 items per witness.

How do these numbers compare to other studies? It is difficult to make direct comparisons because there is a good bit of variation across studies, both in terms of the procedures and in the way the results are reported. Wells et al. (1993) did not report a total number of descriptors, but instead reported an itemized breakdown. Lindsay et al. (1994) reported 7.35 items per free-recall description, 6.36 items excluding clothing and jewelry. Thus, initial descriptions in Lindsay et al. were slightly longer than those of the present study, but the number of items was increased in the present study through the prompts on the description forms.

The purpose of the forms was to simulate the kinds of follow-up questions that might be asked by detectives after obtaining initial free-recall descriptions. Thus, our descriptions might be more comparable to descriptions obtained in archival studies. Two such studies, by Kuehn (1974) and Sporer (1996), reported 7.2 and 9.71 descriptors per description, respectively. With clothing details excluded in the Sporer study, the number of descriptors was 6.59. Our descriptions were slightly longer than those reported by Sporer. There may be several reasons for the longer descriptions in the current study. First, there may be some underreporting in the Sporer study; for example, the gender of the perpetrator was not listed in the itemized listing of descriptors. It is unlikely that none of the witnesses was able to report the perpetrator's gender. In addition, only about one fourth of the witnesses reported the perpetrator's race. This suggests that many of these witnesses did not get a very good look at the perpetrator, which would be reasonable if the conditions of observation spanned the range that might occur in real criminal cases. The conditions in the present experiment were designed so that it would be very unlikely that any of the witnesses would not have seen the perpetrator.

A number of studies reported description data by itemizing the number of details per description on a number of categories (age, race, weight, height, hair, skin, face). The totals for this breakdown were 5.69 for the current study, 5.14 for Wells et al. (1993), 4.70 for Sporer (1996), and 5.31 for Lindsay et al. (1994). Again, using this breakdown, the descriptions in the present study appear to be slightly longer than those of other laboratory and archival studies. There are probably two reasons for this: unlike laboratory studies, which utilize only free recall, the witnesses in the present study were prompted to provide additional details; and unlike the archival studies, the conditions of observation were probably better than those represented by the wider range of criminal investigations.

Confidence Analyses

Confidence. Table 2 reports the results for the confidence ratings. A 2 × 2 × 2 ANOVA showed that prospective confidence was higher than retrospective confidence, $F(1, 113) = 39.71$, $p < .0001$, and confidence was higher in perp-present than for perp-absent lineups, $F(1, 113) = 5.06$, $p < .05$. There was no difference between suspect-matched and description-matched lineups ($F < 1$), and none of the interactions approached statistical significance.

Confidence and Accuracy. Among participants who initially made a selection from the lineup (choosers), the correlation was not reliable between *prospective* confidence and accuracy across all conditions, $r(N = 73) = .04$, ns. However, among choosers there was a reliable positive correlation between *retrospective* confidence and accuracy across all conditions, $r(N = 73) = .33$, $p < .01$. The prospective confi-

Table 2. Average Confidence Judgments for Experiment 1

	Perpetrator present		Perpetrator absent	
	Suspect matched	Description matched	Suspect matched	Description matched
Prospective	3.89	3.89	3.65	3.73
Retrospective	3.31	3.31	2.84	2.78

Note: Average confidence ratings were reported by participants before they viewed lineups (prospective) and after their identification responses (retrospective).

dence judgment made prior to viewing the lineup was a poor predictor of a participant's accuracy in the identification task 2 weeks later. However, confidence judgments reported after making a selection were a better predictor of identification accuracy. This result is consistent with the results of a meta-analysis by Cutler and Penrod (1989) that directly compared pre- versus postidentification confidence and accuracy. These researchers found preidentification confidence (prospective) to be a significantly worse predictor of confidence than postidentification confidence (retrospective).

Explanations for Present Results

Consistent with predictions by Luus and Wells (1991), the police officers in the present study constructed lineups that showed higher foil-to-suspect similarity in suspect-matched lineups than in description-matched lineups. However, using these lineups, we did not replicate the advantage for description-matched lineups shown by Wells et al. (1993) or the trend shown by Lindsay et al. (1994). Wells et al., in particular, showed a large advantage in correct identification rates for description-matched lineups, whereas the present results showed the correct identification rates to be dead even at 53.1%. In fact, we found no significant differences between the suspect-matched method and the description-matched method with regard to correct identification rates, false identification rates, foil identification rates, ‘‘don't know'' responses, and perpetrator not in the lineup responses. These results are contrary to both prior predictions and empirical data (Luus & Wells, 1991; Navon, 1992; Wells et al., 1993).

Given the predictions, the similarity ratings, and the prior experimental results favoring description-matched lineups, the equivalence between suspect-matched and description-matched lineups requires explanation. First, the equivalence in correct identification rates cannot be dismissed as due to low statistical power. The results cannot be characterized as an advantage for description-matched lineups that did not reach statistical significance, because the correct ID rates were identical (53.1%) and the false ID rates, although very low, favored the suspect-matched lineups.

One might question, however, whether the no-difference conclusion is based on too few observations. To address this possibility, we calculated the number of subjects required to obtain a significant difference in correct identification rates with a moderate effect size (ϕ =.3). This calculation (based on power = .80) showed an N of 88, 24 more data points than the 64 on which the results are based. However, it is unlikely that if we had continued to collect data, we would have shown a reliable difference between suspect-matched and description-matched conditions. If we had additional data from 24 participants (12 in each condition of the perp-present lineups), a reliable advantage for description-matched lineups would be obtained only if none of the 12 participants in the suspect-matched lineups picked the suspect and at least 10 of 12 participants in the description-matched condition picked the suspect. The only identification result that showed even a trend in the direction of a description-matched lineup advantage was the slightly lower foil identification rate. However, this result also cannot be attributed to low statistical power; even if the number of participants in the experiment were increased by a factor of 4 (N = 512), the difference in proportions for foil identifications would not reach statistical significance by a chi-square test.

The question remains as to whether the difference between our results and those of Wells et al. (1993) are due to differences in how the lineups were constructed or due to other differences between the two studies. Although both studies used a staged crime procedure, Experiment 1 presented a crime that would involve more contact between witnesses and the thief, and also had a longer retention interval. In the procedure used by Wells et al., the crime occurred more suddenly and ended more quickly, and the identification lineups were presented within minutes in the same session.

The rationale in Experiment 1 was to minimize the cases in which witness errors would be due to a failure to observe rather than a failure of memory over time. The reason for this is that witnesses who indicate that they did not see the perpetrator are probably unlikely to be contacted later for the purposes of identification. To include such witnesses in an experimental study could provide an inappropriately low estimate of accuracy.

Could these differences underlie the differences in the identification results of Experiment 1 compared to Wells et al.? Presumably, the differences in the crime and the retention interval would produce differences in the overall level of performance. The fact that, relative to Wells et al., (1993), Experiment 1 showed higher correct identification rates for suspect-matched lineups, but lower correct identification rates for description-matched lineups suggests that the main difference between the two studies is in the composition of the lineups.

The results of both experiments (and those of Experiment 2, to be presented later) are plotted in Fig. 2 for comparison. From the figure, the differences for the description-matched conditions appear minimal, whereas the patterns of results are quite different for the suspect-matched conditions. Although the similarity in results does not necessarily mean that description-matched lineups were similarly composed in the two studies, it does seem clear that the biggest differences are in the suspect-matched lineups. We suggest that the foils in the suspect-matched lineups were much more similar to the suspect in the Wells et al. (1993) study than in Experiment 1. The supporting evidence for this claim is as follows.

For the perp-present lineups, 25% of the witnesses in Experiment 1 picked a foil, whereas 43% picked a foil in Wells et al. (1993). In addition, for the perp-

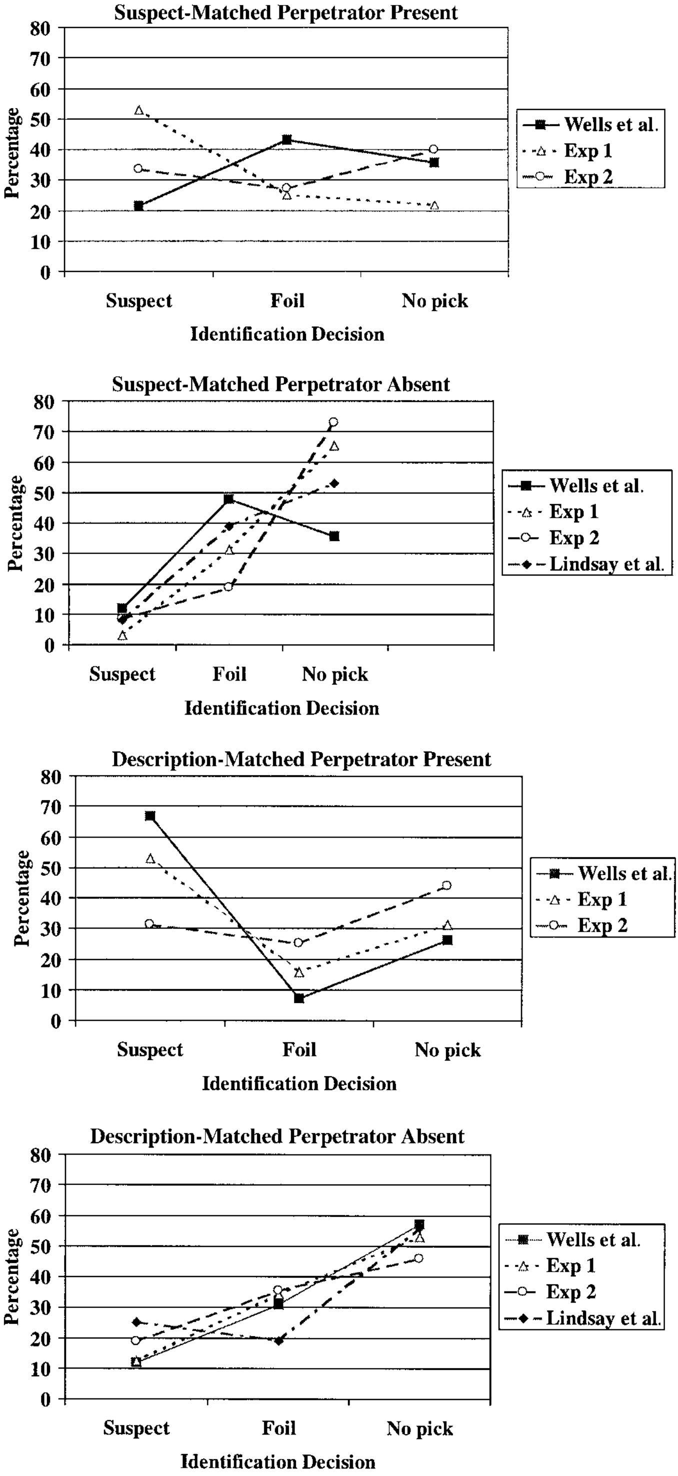


Fig. 2. Comparison of results for experiments 1 and 2 and results of Wells et al. (1993) and Lindsay et al. (1994) (perp-present results for Lindsay et al. were not available).

present lineups, the conditional probability of picking the suspect given that anyone was picked was .68 in Experiment 1, but only .33 for Wells et al. These results strongly suggest that there was much more overlap between suspect and foils in the Wells et al. study than in Experiment 1.

In addition, Wells et al. (1993) specifically stated that their foils were selected to be ‘‘the five photos that *most* resembled the suspect's photo’’ (p. 838, italics added). We did not instruct detectives as to how similar the foils should be to the suspect because we wanted to find out how they would normally pick foils, with the only constraint being that they base their selections on their match to the suspect. Presumably, the officers used their own judgment in determining how similar ‘‘similar’’ should be, in accordance with their usual procedures.

Our results suggest that the police officers in this study did not pick foil photos that *most* resembled the suspect. The similarity rating data do indicate that foil-to-suspect similarity was higher for suspect-matched than for description-matched lineups; however, this similarity difference did not translate into lower correct identification rates in suspect-matched lineups. The combination of these results suggests that overmatching is a potential problem, but reduced identification rates are not a necessary consequence of that problem.

Still unclear is the deviation from predictions regarding false identifications. False identification rates were very low, less than 8% across both suspect-matched and description-matched lineups. We left it up to the officers to pick the suspect in perp-absent lineups, assuming that they would use the same standards of similarity for picking the (innocent) suspect that they would for picking the foils. The results suggest that the innocent suspects they picked were not that similar to the actual perpetrators. We asked the witnesses who did not make a pick to pick the person they would pick if they had to pick someone. When these results are combined with the results of witnesses who picked without such prompting, we should, according to Navon (1992), find that the innocent suspect is picked from suspect-matched lineups more than would be expected by chance. We found no support for this prediction: 6 of 32 witnesses, or 18.8%, picked the innocent suspect, only slightly different from the 16.7% that one would expect by chance.

Clearly, our results are at odds with predictions and previous results that favor description-matched foil selection over suspect-matched foil selection. The differences in the results between Experiment 1 and Wells et al. (1993) appear to be due to the composition of the lineups. Experiment 1 was designed to closely simulate police procedures, and it may be that some aspects of these procedures underlie the lineup differences. Three of these are discussed below.

First, the lineups may be different because police officers have different strategies for picking foils than do researchers. Good research requires strong experimental manipulations and consistency within experimental conditions. These are not necessarily the goals of police officers in creating lineups. Assuming that officers believe they have the right person (i.e., that the suspect and perpetrator are one and the same), a reasonable strategy for lineup construction is to select foils to maximize the likelihood of selecting the suspect while ensuring that the lineup will stand up in court. Thus, the differences in composition may be due to different strategies used by the people making up the lineups. A second factor is that police officers have a much larger and much broader range of photographs than are used in most eyewitness memory experiments. Many experiments select foils from a pool composed entirely of photographs of college students, most of whom are between the ages of 18 and 22 years. Starting off with such a homogeneous pool may sharpen and focus the standard of similarity. Both of these points are addressed in Experiment 2, in which untrained undergraduates created the lineups from a pool of college yearbook photographs.

Finally, in Experiment 1 we followed free recall of descriptions with a form that would prompt witnesses to report additional details. We did this because of discussions with police detectives who indicated that it was standard police procedure to ask follow-up questions after witnesses give descriptions. This point is also echoed in a comment by Lindsay et al. (1994), who said that foil selection based on free recall descriptions only would be ‘‘questionable at best’’ (p. 532). One possible result of this follow-up procedure was that our descriptions appeared to be slightly longer than those reported elsewhere. The longer descriptions may have allowed the detectives who created description-matched lineups to commit the same kind of overmatching errors that Luus and Wells (1991) argued would occur for suspect-matched lineups. This possibility is addressed in Experiment 2 by not using the follow-up forms after free recall.

EXPERIMENT 2

As in Experiment 1, suspect-matched and description-matched procedures were used to construct both perp-present and perp-absent lineups. Otherwise, the procedures in Experiment 2 were quite different from those in Experiment 1. Candid yearbook photographs were used as target stimuli rather than a staged-crime procedure. The lineups were constructed by other college students, rather than by police officers; and the pool of photographs was more homogeneous, consisting of college yearbook graduation photographs. Furthermore, free-recall descriptions were used in description-matched lineups instead of standardized description forms. Experiment 2 also used a within-subjects design; each participant was presented with four lineups, completing the 2 × 2 design (perp-present and perp-absent, suspect- and description-matched lineups). The purpose in using the within-subjects design is to minimize the chances that real differences between suspect-matched and description-matched lineups will go undetected.

Method

Participants

A total of 148 people participated, either in partial fulfillment of an introductory psychology course requirement or for cash payment. Forty-eight people participated as ‘‘witnesses,’’ viewing the target photographs, and then returning 1 week later to make identification responses. Another group of 48 participants created the lineups.

The remaining 52 people provided foil-to-suspect similarity ratings for each of the lineups.

Materials

Four candid photographs from a college yearbook were used as the target stimuli. Each candid photograph showed one person in a natural environment (dorm room, outdoor park area, etc.). The four people in these photographs looked quite different, so there would be little confusion between them. They were a dark-haired, brown-eyed male; a dark-haired, brown-eyed female; a light-haired, blue-eyed male; and a light-haired, blue-eyed female. These people constituted the ‘‘perpetrators.’’ Four people who looked similar to the perpetrators were selected to be used as ‘‘innocent’’ suspects. A pool of 140 senior-year photographs was used to create lineups.

Procedure

Session 1

Forty-eight people participated individually as witnesses. Each witness was shown four photographs, one at a time, for 15 sec each. After each photograph was removed, the witness solved arithmetic problems for 1 min as a distractor task, and then was given up to 5 min to write a description of the person in the photograph on a blank notecard. Each participant completed this task four times (once for each perpetrator). The order of presentation of the photographs was counterbalanced across participants so that each target person appeared in each of the four order positions. After observing and describing the four perpetrators, the witness was dismissed and asked to return 1 week later; the purpose of the second session was not revealed at that time.

Photo-Lineup Construction. A separate group of 48 participants created lineups using suspect-matched and description-matched methods. Each participant created four lineups, one for each of the four conditions: perp-present, suspect-matched; perp-absent, suspect-matched; perp-present, description-matched; and perp-absent, description-matched. The order in which the four different lineups was created was counterbalanced. For all perp-absent lineups, the innocent suspect was selected by having a separate group of participants compare and rate each perpetrator to the pool of photographs according to similarity. The person who was rated the highest in similarity to the perpetrator was selected as the innocent suspect (one each per perpetrator). Four rather than five foils were selected for each lineup. The purpose of the smaller lineup size was so that people would not compromise their standards for similarity in order to fill all positions in the lineup.

For the suspect-matched lineups, each participant was given the head-and-shoulder yearbook photograph of either the perpetrator or the innocent suspect. Four foils were then selected based on the match to the suspect photograph.

For the description-matched lineups, participants received only the description of the perpetrator and not the photograph itself. Four foils were chosen based on the match to the description. For each of these lineups, the experimenter later added the suspect.

Table 3. Identification Responses for Experiment 2

	Perpetrator present		Perpetrator absent	
	Suspect matched (%)	Description matched (%)	Suspect matched (%)	Description matched (%)
Suspect	33.3	31.3	8.3	18.8
Foil	27.1	25.0	18.8	35.4
No pick	39.6	43.8	72.9	45.8

Note: A comparison of suspect-matched and description-matched conditions for suspect identifications, foil picks, and no picks.

Session 2

Participants who had originally given descriptions based on the candid photographs returned individually to look at lineups. Prior to viewing the lineups, the witness was read an admonition similar to that used in Experiment 1. For each of the four lineups, the witness either made a selection or did not pick,[8] and then was asked to rate his or her confidence. Each lineup corresponded to one of the perpetrators described in session 1 and the presentation order was counterbalanced.

Similarity Ratings. Fifty-two people from the same participant pool, who did not see the candid photographs, rated the similarity of each foil to the suspect in each lineup. The ratings were made on a scale of 1–7 for the lowest to highest similarity. As in Experiment 1, each person rated both description-matched and suspect-matched lineups in either the perp-present ($N = 27$) or perp-absent ($N = 25$) condition.

Results

Similarity Ratings

As in Experiment 1, average ratings were computed for each individual lineup and then averaged across all lineups within each condition. For perp-present lineups, the mean ratings were 2.57 ($SD = .764$) for suspect-matched and 2.10 ($SD = .615$) for description-matched. For perp-absent lineups, averages were 2.84 ($SD = .793$) and 2.27 ($SD = .645$), respectively. Again, similarities between suspect and foil were higher in suspect-matched than in description-matched lineups for both perp-present, $t(26) = 12.188$, $p < .001$, and perp-absent, $t(24) = 11.355$, $p < .001$, lineups, with large effect sizes of .922 and .918.[9]

Identification Responses

The response probabilities are given in Table 3 for suspect- and description-matched lineups for both perp-present and perp-absent conditions. Visual inspection of the table shows that the patterns of responses for suspect- and description-matched lineups were again very similar when the perpetrator was in the lineup.

[8] For Experiment 2, we did not subdivide the no-pick responses into ‘‘don’t know’’ and ‘‘reject’’ responses.

[9] The Wilcoxon Signed Ranks Test showed significant differences for perp-present, $z(26) = -4.541$, $p < .001$, and perp-absent, $z(24) = -4.373$, $p < .001$, conditions.

All of the response probabilities were within 4.5%. For perp-absent lineups, suspect-matched and description-matched lineups showed somewhat different patterns of responses.

The results were statistically analyzed using *t*-tests (because of the within-subjects design[10]) comparing suspect-matched and description-matched lineups for picking the suspect, picking a foil, and picking no one from the lineup. There were no differences in response rates for picking the suspect for perpetrator present or absent lineups, $t(48) = .227$, ns; $t(48) = 1.53$, ns, respectively. In order to make this difference reliable ($p < .05$) for the perp-present condition, 10 additional subjects would have to select the suspect in the suspect-matched condition and never pick the suspect in the description-matched condition. There were also no significant differences for picking a foil, or declining to pick (all *t*s < 1) for perp-present lineups. For perp-absent lineups, there were no differences in response rates for picking the innocent suspect, $t(47) = 1.528$, ns, or picking a foil, $t(47) = 1.832$, ns. However, witnesses were more likely to decline to pick for suspect-matched (72.9%) than for description-matched (45.8%) lineups, $t(47) = 3.098$, $p < .005$.

Conditional probabilities of false identifications in perp-absent lineups were .308 and .346 for suspect-matched and description-matched lineups, respectively. For these five-person lineups, chance performance would be .20. The suspect-matched probability did not differ significantly from chance ($z = .971$), but the description-matched probability was marginally different ($z = 1.863$, $p < .06$). The difference in the outcomes of the statistical tests is due to the larger number of observations in the description-matched condition because more witnesses made a pick.

The identification results are plotted in Fig. 2 with those from Experiment 1 and previous results from Wells et al. (1993) and Lindsay et al. (1994). The same pattern emerges that was noted earlier: The results are very consistent across experiments for the description-matched lineups, but not for suspect-matched lineups. This variability in the suspect-matched lineups is consistent with Luus and Wells'(1991) criticism of suspect-matching—that there is no rule for how similar is similar enough.

Functional Size. Functional size was calculated using the same method as for Experiment 1. This analysis showed functional sizes of 2.88 for description-matched lineups and 3.25 for suspect-matched lineups. For both conditions, the functional sizes were lower than would be expected from a fair lineup; however, the differences between conditions were very small.

Witness Descriptions

Descriptions. Witness descriptions contained an average of 8.74 items; however, when irrelevant information was excluded (i.e., clothing and jewelry), this number decreased to an average of 6.13 items. Our numbers are similar to those reported by Sporer (1996) and Lindsay et al. (1994), who found on average 6.59

[10] The within-subjects design precludes the use of chi-square analyses. Although the use of *t*-tests may not be ideal for nominal-level data, we wanted to use as powerful a statistical test as possible to minimize the likelihood of missing real differences between the critical conditions.

and 6.36 items, respectively. An analysis of the subset of items (i.e., age, race, weight, hair, skin, and face) showed the results of the present study to be slightly lower than in other studies. For Experiment 2, Wells et al. (1993), Sporer (1996), and Lindsay et al. (1994), in that order, the numbers are as follows: 4.48, 5.14, 4.70, and 5.31.

Confidence Analyses

Confidence. Table 4 reports the data for the confidence ratings. There was no difference in the average confidence for perp-present and perp-absent lineups, no difference between suspect- and description-matched lineups, and no interaction (all Fs < 1).

Confidence and Accuracy. Among participants who made a selection (1–5) from the lineup, there was no reliable correlation between confidence and accuracy across all conditions, $r(N = 192) = .14$, ns. Therefore, the confidence judgments given after identification decisions were made were not a good predictor of a participant's accuracy.

Summary of Experiment 2 Results

Extended discussion of Experiment 2 results is reserved for the General Discussion section. Here, we summarize and comment only briefly on the results. The basic pattern of results from Experiment 1 was replicated in Experiment 2: Similarity ratings indicated higher foil-to-suspect similarity for suspect-matched lineups than for description-matched lineups. However, suspect-matched and description-matched lineups did not differ in correct or false identifications. For perp-absent lineups, suspect-matched lineups showed more correct no-pick responses than did description-matched lineups.

The results are consistent despite a number of substantive differences between Experiments 1 and 2. We had speculated that the results of Experiment 1 might have been due to certain aspects of police procedures that were simulated in Experiment 1 (i.e., particular strategies; a larger, more heterogeneous pool of photographs; or follow-up questions that would add more detail to witness descriptions). However, the very same pattern of results was obtained with lineups created by undergraduates who selected foils from college yearbooks, based on shorter free recall descriptions. Thus, there is no evidence that the results of Experiment 1 were obtained as a result of any of these factors.

Table 4. Average Confidence Judgments for Experiment 2

	Foil selection method	
	Suspect matched	Description matched
Perpetrator present	3.35	3.23
Perpetrator absent	3.52	3.35

Note: Average confidence ratings related to participant identification decisions.

GENERAL DISCUSSION

Two experiments tested the following predictions: (1) Foils selected based on their match to a photograph of the suspect will be more similar to that suspect than foils selected based on their match to a description of the perpetrator. (2) Correct identification rates should be higher for description-matched than for suspect-matched lineups. (3) False identification rates should also be higher for suspect-matched lineups than for description-matched lineups. (4) Conditional false identification rates in suspect-matched lineups should be significantly greater than chance. The first of these predictions was supported by similarity rating data. However, none of the other predictions was supported.

The similarity rating and identification data were remarkably consistent across experiments, despite a number of substantive differences. This consistency suggests that the results are not due to peculiar strategies that might be used by police officers or the large, heterogeneous pool of mugshots used to select foils. Likewise, the identification results were consistent regardless of whether longer descriptions were obtained through follow-up prompting (Experiment 1) or whether the descriptions were fairly short (Experiment 2). Also, the pattern of identification results does not appear to be unique to forensically relevant simulations, as it occurred when participants examined still photographs with the intention of writing a description of the person in the photograph. In addition, the pattern held up when participants were presented with a single lineup or with four different lineups. The results were consistent when lineups were created by experienced police officers or by untrained college students.

Correct Identifications

Correct identification rates were identical in Experiment 1 and differed by the response of one witness in Experiment 2. From these results, we conclude that there is nothing inherent in the suspect-matched method of foil selection that *by necessity* produces low rates of correct identification. Nonetheless, as was clearly demonstrated by Wells et al. (1993), if foils *are* selected to be very similar to the suspect, the correct identification rate can be drastically reduced.

False Identifications

The present results provide no evidence that suspect-matched lineups are unfairly biased against innocent suspects. False identification rates were low, and conditional false identification rates did not deviate from chance.

A surprising result of the present experiments is that there was a consistent trend toward more bias in description-matched lineups than in suspect-matched lineups. False identification rates were higher for description-matched lineups in both experiments, although the rates were too low to interpret statistically. However, in both experiments the conditional false identification rates were higher for description-matched lineups, and for Experiment 2 the deviation from chance was marginally significant.

These trends might seem unimportant; however, this same pattern of conditional probabilities was shown in both previous experiments. Lindsay et al. (1994)

showed conditional false identification rates of .176 and .563 for suspect-matched and description-matched lineups, respectively. Wells et al. (1993) showed these probabilities to be .200 and .278. None of these probabilities differed significantly from chance (all *z*s < 1), with the exception of the description-matched condition in the Lindsay et al. experiment ($z = 4.249$, $p < .001$).

The statistical analyses preclude any strong conclusions; however, the pattern of results is exactly the same in all four experiments, as shown in Fig. 3. All four experiments show higher conditional false identification rates for description-matched than for suspect-matched lineups. Moreover, the conditional probabilities are close to chance for the suspect-matched lineups, and either slightly higher or significantly higher than chance for description-matched lineups.

What conclusions might be drawn from these conditional probabilities? First, suspect-matched lineups do not appear to be biased against innocent suspects. The consistency in the results suggests that description-matched lineups may be biased, particularly in the Lindsay et al. (1994) study. However, this conclusion may be unwarranted. Although the pattern is consistent, only the Lindsay et al. results were unambiguously reliable, and those results may be due to the specifics of how the description-matched lineups were constructed for that study rather than to any inherent flaws in the description-matching procedure. Specifically, Lindsay et al. used a ''modal'' description to select foils for description-matched lineups instead of using individual witness descriptions. As a result, several witness descriptions may not have matched the modal description, and thus the foils may have not been very similar to a given witness's memory of the perpetrator. For example, in their study, 80% of witnesses agreed that the perpetrator had ''dark hair''; however, if

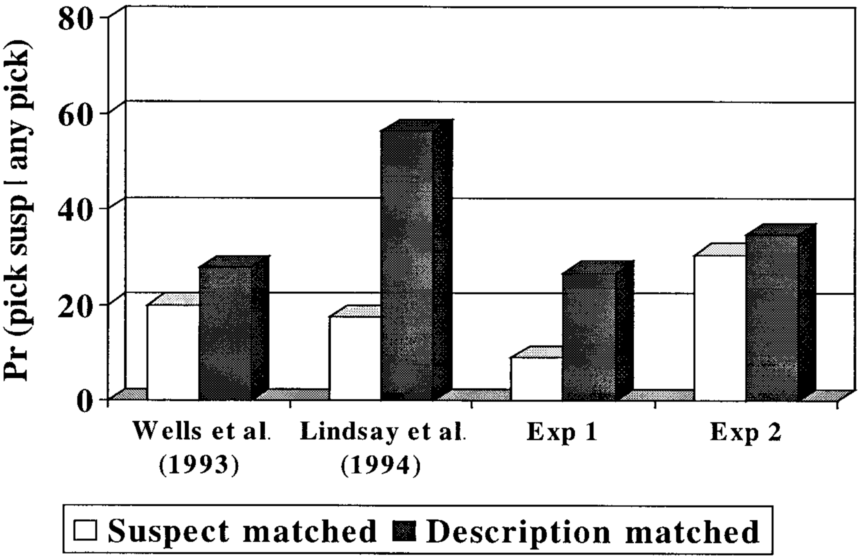


Fig. 3. Conditional probabilities of falsely identifying the innocent suspect given that the witness makes any pick, for experiments 1 and 2, Wells et al. (1993), and Lindsay et al. (1994).

a given witness reported brown hair instead, it is possible that none of the foils would match the witness's memory, allowing many foils to be easily rejected. The clear implication of this for law enforcement is that when there are several witnesses, description-matched lineups *must* be custom-fit to the description given by each witness, rather than based on a composite or averaged description. Still, the consistency of the overall pattern calls for additional research on this issue.

It is puzzling that Navon's (1992) predictions, which are based on very simple and general assumptions about similarity, have not been confirmed. If the innocent suspect is selected based on his or her match to a description of the perpetrator, but the foils are matched only through second-order similarity to the suspect, how can the innocent suspect *not* be the best match (on the average) to the witness's memory, compared to any of the foils? The failure to confirm this prediction suggests that witnesses may be using more complex decision strategies when selecting from perp-absent lineups.

Correct Rejections

Experiments 1 and 2 show that correct rejections were higher for suspect-matched than description-matched lineups. The explanation is quite simple: The suspect-matched method selects foils who are similar to the wrong person, and therefore they present the witness with a lineup in which no one looks quite like the actual perpetrator. This result has important implications. For the innocent suspect, whether the witness picks a foil or picks no one is of little consequence. However, it is a different matter for the police conducting the investigation. Consider a situation in which a witness is shown two lineups (perhaps days or months apart). The first lineup contains an innocent suspect who is later exonerated, and the second lineup contains the true perpetrator, who is correctly identified by the witness. The witness's credibility would likely be bolstered by having correctly rejected the first lineup, but would be undermined by having made a false pick on the first lineup, even of a foil.

Similarity and Identification

Identification and Foil-to-Suspect Similarity

Both experiments showed a dissociation between the identification results and the foil-to-suspect similarity ratings. This is particularly surprising in the perp-present conditions because the suspect *is* the perpetrator. How can the foils in suspect-matched lineups be more similar to the perpetrator than the foils in the description-matched lineups without a proportional decrease in correct identification rates? Although this result remains a puzzle, it is clear that the relationship between similarity and identification responses is not straightforward. It may be that the similarity difference becomes relevant only if the overall similarity is very high. It might also be that the similarity differences, although reliably different, were not large enough to produce differences in identification results.

A complete understanding of the relationship between similarity and identification, including the similarity–identification dissociation, may require extensive theoretical development.

Possible Effects of Delay?

In the present research the lineups were presented 1 week (Experiment 2) or 2 weeks (Experiment 1) after seeing the target person, whereas in previous research (Lindsay et al., 1994; Wells et al., 1993) the identifications were made minutes later within the same experimental session. This leaves open the question as to whether the present results are somehow a result of the longer retention interval. This can only be the case if the time delay interacts with foil selection method in such a way as to produce lower performance primarily for the description-matched lineups so as to eliminate the description-matched lineup advantage that would have otherwise been found.

We cannot rule out this possibility because there is no empirical comparison of short and long retention intervals combined with a comparison of foil selection methods. However, we know of no mechanism that would produce the kind of interaction that would implicate retention interval as a mediator in the present results. Also, if the advantage for description-matched lineups were eliminated by the longer delay, this would mean that the generality of the advantage is limited only to those cases for which the lineup is presented within 1 week.

How Similar Is Similar Enough?

This question is still not answered by suspect-matched lineups. The results of Experiment 1 suggest that the police officers who made lineups did not pick foils that were highly similar to the suspect for either the suspect-matched or description-matched lineups. Although this moderate level of similarity produced relatively high accuracy for perp-present lineups, the concern is that it might also lead to high false identification rates. The low levels of false identification shown in experiments 1 and 2 should not assuage these concerns, for it may be that misidentifications were avoided because not only were the foils not terribly similar to the innocent suspect, but the innocent suspect was also not terribly similar to the true thief. To some extent, this concern was addressed in Experiment 2, in which the innocent suspect was preselected to be very similar to the perpetrator.

‘‘Similar enough’’ may depend in part on how similar the innocent suspect is to the perpetrator. An innocent suspect who is very similar to the actual perpetrator would presumably only be protected from false identification by foils who are also very similar to the actual perpetrator. Given this logic, it would seem that in order to answer the ‘‘similar enough’’ question, we must first have some information about the similarity between the perpetrator and the innocent suspect who is to be protected from false identification. The relevant questions are (1) In what proportion of real criminal investigations where an innocent person is presented in a lineup will that innocent suspect be a dead-ringer for the real perpetrator? (2) In what proportion of those cases will the innocent suspect not be very similar to the innocent suspect? In other words, we need to know what the expected similarity is between innocent suspects and the real perpetrators in real criminal cases. Of course, in order to know these expected similarities, one would need data from cases in which an innocent suspect is arrested, later exonerated, and the true perpetrator is later known. To our knowledge, no systematic similarity analyses have been conducted for such cases.

Degree of Similarity or Kind of Similarity?

Although suspect-matched foil selection can result in gratuitous similarity between the suspect and foils (Wells et al., 1993), our results showed no evidence that this is a necessary outcome. There may, nonetheless, be important differences between suspect-matched and description-matched lineups. The key issue may not be about the degree of similarity, but rather the target of the similarity and the kinds of features that drive the match. Many theorists hold that faces are processed not in terms of individual features, but rather in terms of high-level, holistic, and configural features. Verbal descriptions of faces, however, tend to be presented as lists of features (Wells & Turtle, 1987). From this, one might expect that suspect matching allows matching on pictoral and configural features, whereas description matching matches on individual and separate features. How this might affect identification accuracy remains for future research.

Consistency of Implementation

Luus and Wells (1991) argued that one problem of suspect-matching foil selection is that there is no clear answer to the ‘‘similar enough’’ question. This problem, they argued, can be solved by description matching because it is clear which features to match and when to stop. Based on this, one might expect that there should be greater variance in creating suspect-matched lineups than description-matched lineups, and from this, one would predict less variance in identification performance in description-matched lineups than in suspect-matched lineups. A comparison of the four relevant experiments supports this line of reasoning.

The results of experiments 1 and 2 are plotted in Fig. 2, along with results from Wells et al. (1993) and Lindsay et al. (1994). It is clear from the figure that the experiments showed very different results for suspect-matched lineups, but strikingly similar results for description-matched lineups. These results suggest that the implementation of description-matched lineups may be less subjective than the implementation of suspect-matched lineups. The consequence of this is that it may reduce the likelihood of producing a lineup with foils that are too similar or not similar enough.

CONCLUSIONS AND FUTURE RESEARCH

Two experiments showed no difference between suspect-matched and description-matched lineups for perp-present lineups and showed an advantage in correct rejections in perp-absent lineups for suspect-matched foils. These results are contrary to predictions (Luus & Wells, 1991; Navon, 1992) and previous results (Wells et al., 1993; Lindsay et al., 1994). Furthermore, our results provide a challenge for recommendations favoring the description-matched method made by Wells et al. (1998; Rule 3) and may suggest that such recommendations are premature. The results by Wells et al. (1993) provide a clear and compelling demonstration that suspect-matched lineups with foils that are very similar to the suspect can produce very low rates of correct identification. Moreover, because the suspect-matched method does not provide a clear ‘‘similar-enough’’ rule for selecting foils, there is no mechanism to guarantee that such lineups will not be created. What the present results show, however, is that such lineups are not an unavoidable consequence of suspect-matched foil selection, and indeed, such lineups may occur quite infrequently. In addition, the present results showed an advantage for suspect-matched lineups over description-matched lineups in terms of correctly rejecting perp-absent lineups, a point that should not be overlooked in considering the relative merits of the two procedures. However, the present results should not be taken as an endorsement for suspect-matched lineups. Rather, the present results and the analyses of previous results suggest that the issue of picking foils for lineups is complex, and additional research is needed.

The present results point to a complex relationship between similarity and eyewitness identification. Future research should investigate these similarity relationships. A central question in the present research is how similar the foils should be to the suspect in a lineup. Assuming that foil-to-suspect similarity has a role in protecting innocent suspects, the answer to the ‘‘similar enough’’ question likely depends on the similarity between the innocent suspect and the true perpetrator. In addition, the present results showing a dissociation between identification responses and foil-to-suspect similarity ratings suggest that comparisons of similarity ratings and identification responses may offer some new insights into the cognitive processes underlying identification.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation grant HS 97-056 to S.E.C. We wish to thank the San Bernardino County Sheriff’s Department for helping with the lineup construction, and we are particularly grateful to Sheriff Gary Penrod, Sheriff’s Deputy Chief Mike Cardwell, and Captain Mike Ingram for consulting with us on this project. Thanks also to Doug Wells and Eric Caston for their performances as thieves, and Cecelia O’Drobinak and Dominique Hysmith for their invaluable help with data collection. We would like to thank Gary Wells and Rod Lindsay for insightful discussions and providing us with data. Finally, we appreciate the advice of an anonymous reviewer and Richard Wiener, who recommended the addition of the similarity rating analyses.

REFERENCES

Brooks, N. (1983). *Police guidelines: Pretrial eyewitness identification procedures.* Ottawa: Minister of Supply and Services.

Clark, S. E. (1999). *WITNESS: A computer simulation model of eyewitness identification.* Presented at the third biennial meeting of the Society for Applied Research in Memory and Cognition, Boulder, Colorado.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*(1), 37–60.

Cutler, B. L., & Penrod, S. D. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology, 74*(4), 650–652.

Greene, E. (1988). Judge's instructions on eyewitness testimony: Evaluation and revision. *Journal of Applied Social Psychology, 18,* 252–276.

Helmreich, R., & Stapp, J. (1974). Short forms of the Texas Social Behavior Inventory (TSBI), an objective measure of self-esteem. *Bulletin of the Psychonomic Society, 4*(5A), 473–475.

Huff, C. R., Rattner, A., & Sagarin, E. (1986). Guilty until proven innocent: Wrongful conviction and public policy. *Crime & Delinquency, 32*(4), 518–544.

Kuehn, L. L. (1974). Looking down a gun barrel: Person perception and violent crime. *Perceptual & Motor Skills, 39*(3), 1159–1164.

Langer, E. J., & Abelson, R. P. (1972). The semantics of asking a favor: How to succeed in getting help without really dying. *Journal of Personality and Social Psychology, 24*(1), 26–32.

Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law and Human Behavior, 18,* 527–541.

Loftus, E. F. (1974, December). Reconstructing memory: The incredible eyewitness. *Psychology Today,* 117–119.

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15,* 43–57.

Navon, D. (1992). Selection of lineup foils by similarity to suspect is likely to misfire. *Law and Human Behavior, 16,* 575–593.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Sporer, S. L. (1996). Psychological aspects of person descriptions. In S. L. Sporer, R. S. Malpass, & G. Koehnken (Eds.), *Psychological issues in eyewitness identification.* Mahwah, NJ: Erlbaum.

Wells, G. L., (1993). What do we know about eyewitness identification? *American Psychologist, 48*(5), 553–571.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78,* 835–844.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M, & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647.

Wells, G. L., & Turtle, J. W. (1987). What is the best way to encode faces? In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 163–168). New York: Wiley.

Wogalter, M. S., Malpass, R. S., & Berger, M. A. (1993). How do police officers construct lineups: A national survey. *Proceedings of the Human Factors and Ergonomics Society, 37,* 640–644.