Suspect-filler similarity in eyewitness lineups: A literature review and a novel methodology

Ryan J. Fitzgerald, Chris Oriet, and Heather L. Price

University of Regina

Abstract

Eyewitness lineups typically contain a suspect (guilty or innocent) and fillers (known innocents). The degree to which fillers should resemble the suspect is a complex issue that has yet to be resolved. Previously, researchers have voiced concern that eyewitnesses would be unable to identify their target from a lineup containing highly similar fillers; however, our literature review suggests highly similar fillers have only rarely been shown to have this effect. To further examine the effect of highly similar fillers on lineup responses, we used morphing software to create fillers of moderately high and very high similarity to the suspect. When the culprit was in the lineup, a higher correct identification rate was observed in moderately high similarity lineups than in very high similarity lineups. When the culprit was absent, similarity did not yield a significant effect on innocent suspect misidentification rates. However, the correct rejection rate in the moderately high similarity lineup was 20% higher than in the very high similarity lineup. When choosing rates were controlled by calculating identification probabilities for only those who made a selection from the lineup, culprit identification rates as well as innocent suspect misidentification rates were significantly higher in the moderately high similarity lineup than in the very high similarity lineup. Thus, very high similarity fillers yielded costs and benefits. Although our research suggests that selecting the most similar fillers available may adversely affect correct identification rates, we recommend additional research using fillers obtained from police databases to corroborate our findings.

Keywords: Eyewitness, Identification, Lineup, Similarity, Filler

Suspect-filler similarity in eyewitness lineups: A literature review and a novel methodology

When constructing a lineup for eyewitness identification, investigators have been advised to ensure fillers – lineup members who are known to be innocent – do not bear too strong a resemblance to the lineup member suspected of the crime. The not-too-similar recommendation is grounded in the idea that selecting fillers who closely match the suspect's appearance will essentially result in a lineup of 'clones' and make correct identifications too difficult (Wells & Luus, 1990). A multitude of sources have endorsed the not-too-similar recommendation (e.g., Brewer & Palmer, 2010; Malpass, Tredoux, & McQuiston-Surrett, 2007; Wells et al., 1998), so one might expect a relatively large database of rigorous empirical research demonstrating a negative effect of fillers who too strongly resemble the suspect. However, our literature review suggests empirical support for this recommendation is sparse. Given the paucity of empirical research demonstrating the utility of the not-too-similar recommendation, one might ask why eyewitness scientists have accepted it so readily. On other lineup identification issues, researchers have required substantial evidence before uniformly endorsing a procedure. For example, the question of whether lineup members should be presented simultaneously or sequentially has been the subject of lively debate (e.g., Lindsay, Mansour, Beaudry, Leach, & Bertrand, 2009; McQuiston-Surrett, Malpass, & Tredoux, 2006; Steblay, Dysart, & Wells, 2011). We suspect the not-too-similar recommendation has not been subjected to the same degree of rigor as other proposed lineup reforms because the similarity-difficulty relation is both intuitive and well-established in other domains within cognitive psychology.

An eyewitness lineup is ultimately a multiple-choice recognition test. Cognitive psychologists have used strong language to describe similarity's effect on such tests, noting that "we can make any recognition test as difficult as we want simply by making distractors similar to

the correct alternative" (Glass, Holyoak, & Santa, 1979, p. 65). Psychometricians have also emphasized the association between similar distractors and item difficulty (Guttman & Schlesinger, 1967; Smith & Smith, 1988). Even in the eyewitness identification domain, it is true that correct identifications are more likely to occur when fillers are very dissimilar than when they are moderately or highly similar (Fitzgerald, Price, Oriet, & Charman, 2013). If the relation between similarity and difficulty is linear, the idea that witnesses would have a better chance of identifying a target accompanied by moderate relative to high similarity fillers would seem to be a logical extension of this principle. However, the evidence from empirical investigations is remarkably unconvincing.

Defining Similarity

The resemblance between two persons is best conceptualized within the framework of a similarity continuum. However, eyewitness researchers have traditionally conceptualized similarity in categorical terms. Police typically do not have the resources to generate fillers on a continuously defined scale of similarity, so the categorical approach is necessary for formulating lineup construction recommendations that can be practically implemented.

For our literature review, we make a distinction between three categories of lineups: biased, moderate similarity, and high similarity. Biased lineups are those which contain fillers who are highly dissimilar to the suspect or fillers who, for some reason, make the suspect stand out. Moderate similarity lineups are those which contain fillers who match a general description of the target, but have not been closely matched to the suspect's appearance. High similarity lineups contain fillers who have been closely matched to the appearance of the suspect/target. Note that the distinction between categories should be considered in relative terms. For example, researchers have demonstrated greater similarity in the high relative to moderate similarity

lineups, but these 'high' similarity fillers have typically yielded similarity ratings near the midpoint of any given scale (Fitzgerald et al., 2013). Thus, lineups that are categorized as high similarity need not score high on an absolute scale of similarity.

Empirical Research on Lineup Member Similarity

At some level, less similar fillers make correct identifications easier. Numerous empirical investigations reveal that correct identification rates are higher for lineups with very dissimilar fillers relative to lineups with similar fillers (e.g., Carlson, Gronlund, & Clark, 2008; Gronlund, Carlson, Dailey, & Goodsell, 2009; Lindsay, Martin, & Webber, 1994). In other words, witnesses are better able to identify a target from a biased lineup than from a fair lineup; however, this would likely also be true of nonwitnesses who are given nothing more than the target's description (Doob & Kirshenbaum, 1973). Given that the extant literature has consistently shown that correct identifications are less likely to occur in fair relative to biased lineups, we focus our review on the more contentious distinction between lineups containing fillers of moderate and high similarity to the target.

Although a relatively substantial literature on lineup member similarity has been established, the wide variety of methods used to manipulate similarity has made a parsimonious interpretation of the findings elusive. Most commonly, researchers have manipulated similarity through filler selection strategies or by using similarity ratings to guide lineup construction; however, alternative methods can also be found in the literature. Given the potential influence of the type of similarity manipulation on the pattern of identification responses, we have organized our literature review according to the method by which similarity was manipulated. In addition to our primary interest in the effect of high similarity fillers on correct identification of a guilty suspect, we also review their effect on false identification of an innocent suspect.

Filler search procedure: Match-to-description vs. match-to-appearance. Luus and Wells (1991) recommended matching fillers only to the features that were noted in the eyewitness description of the culprit. They hypothesized matching to a description would (a) prevent false identifications by ensuring that all lineup members correspond with the witness's recall of the culprit's appearance and (b) facilitate recognition of the culprit by promoting variation in the lineup members' facial features. The match-to-description strategy was contrasted with a strategy of matching fillers to the suspect's appearance, which Luus and Wells hypothesized would lead to excessive homogeneity in facial features and impede recognition of the culprit. Relative to description-matched fillers, appearance-matched similar fillers were hypothesized to bear a stronger resemblance to the suspect, and this increased similarity was hypothesized to reduce correct identifications and provide no additional protection for innocent suspects.

Wells, Rydell, and Seelau (1993) conducted the first empirical comparison between the match-to-description and match-to-appearance procedures. As Luus and Wells (1991) predicted, the two procedures did not differ in false identifications and the correct identification rate was substantially reduced in appearance-matched lineups. However, subsequent comparisons of the two procedures revealed no significant differences in correct identifications (see Table 1) and in one experiment description-matched fillers increased false identifications (Lindsay et al., 1994). The overview of all comparisons between the match-to-description and match-to-appearance procedures in Table 1 shows the effect on correct identifications has diminished since the time the effect was first reported in the literature (i.e., a decline effect; Schooler, 2011). Furthermore, meta-analytic comparisons of the two procedures suggest that whether fillers are selected by matching to a description or matching to the suspect's appearance has no effect on the extent to

which suspect identifications are diagnostic of guilt or of innocence (Clark & Godfrey, 2009; Clark, Howell, & Davey, 2008).

Subjective similarity ratings. In many of the comparisons between appearance- and description-matched lineups, researchers obtained ratings to measure the similarity between the suspect and the fillers. When such ratings have been obtained, suspect-filler similarity has been consistently higher for appearance-matched lineups than for description-matched lineups. Thus, the differences associated with similarity ratings in these studies can be inferred by examining Table 1 and substituting "moderate similarity" with description-matched and "high similarity" with appearance-matched.

In contrast to indirectly manipulating similarity through filler selection procedures, Brewer and Wells (2006) used subjective ratings to manipulate suspect-filler similarity. Brewer and Wells administered two lineup identification tasks, one for a thief and one for a waiter. In both lineups, some fillers from a high similarity lineup were replaced with fillers who matched a description of the target, but were rated to be of lower similarity to the target. This similarity manipulation yielded mixed results. For the thief lineup, the correct identification rate was higher for the high similarity lineup (.40) than for the moderate similarity lineup (.34). The opposite was true for the waiter lineup, which yielded a higher correct identification rate for the moderate similarity lineup (.66) than for the high similarity lineup (.57). False identifications were comparable across similarity conditions. Given the small and contradictory effects on correct identifications and null effects on false identifications, this manipulation of similarity seems to have had minimal impact on identifications.

Sauer, Brewer, and Weber (2008) used subjective ratings to manipulate similarity in lineups for male and female targets. All lineup members were consistent with a general

description of the target person, but similarity ratings were higher for the high similarity lineup members. In one condition, participants were instructed to respond to the lineup with a traditional identification response (binary condition). For these participants, increasing similarity led to small, nonsignificant reductions in correct identification rates (6% decrease for female lineups; 16% decrease for male lineups). These trends were also observed for participants in another condition who were instructed to rate their confidence that a lineup member was the target rather than actually picking a lineup member. When collapsed across conditions, increasing similarity yielded a small but significant decrease in accuracy for target-present lineups and had virtually no effect on accuracy for target-absent lineups.

More recently, researchers have used subjective ratings to manipulate similarity in lineups for child witnesses (Fitzgerald, Whiting, Therrien, & Price, 2014). When lineup member similarity was increased in Experiment 1, children were significantly less likely to identify the culprit (moderate = .23 vs. high = .07) and also significantly less likely to identify an innocent suspect (moderate = .30 vs. high = .04). In Experiment 2, similarity was manipulated in lineups for child and adult witnesses. For children, the pattern observed in Experiment 1 was replicated in Experiment 2. That is, relative to children in the moderate similarity condition, children in the high similarity condition were significantly less likely to identify the culprit (moderate = .74 vs. high = .48) or misidentify the innocent suspect (moderate = .33 vs. high = .14). Although the high similarity fillers also significantly reduced the adults' innocent suspect misidentification rate (moderate = .28 vs. high = .15), the similarity manipulation had no effect on the adults' correct identification rate (moderate = .76 vs. high = .74). Thus, although child witnesses seem to have difficulty with high similarity lineups, the high similarity fillers only had positive effects on the adult witnesses.

Tredoux's *E'*. Effective size represents the number of lineup members who are plausible alternatives to the suspect (Malpass, 1981). Effective size calculations require data from mock witnesses who are provided a description of the culprit and asked to choose the lineup member who best fits the description (Doob & Kirshenbaum, 1973). Effective size is estimated as the number of lineup members chosen at a rate that differs from chance expectancy. For a 6-member lineup, an effective size of 3 has been proposed to be fair (Brigham, Ready, & Spier, 1990). Tredoux (1998) describes a similar measure, *E'*, which retains the favourable properties of effective size and also utilizes a known sampling distribution. Both effective size and *E'* are positively associated with suspect-filler similarity (Brigham & Brandt, 1992; Brigham et al., 1990; Tredoux, 2002).

Carlson and colleagues (2008) used photos from a mugshot database to construct 'intermediate' (E' = 2.56) and 'fair' (E' = 4.05) lineups. Intermediate lineups contained a mix of fillers who did and did not match the culprit's description. All fillers matched the description in fair lineups. The fair lineups led to a consistently lower rate of innocent suspect misidentifications and although the lineup manipulation did not affect overall correct identification rates (intermediate = .33; fair = .36), an interesting pattern emerged when lineup presentation was taken into account. When presented simultaneously the correct identification rate was higher for intermediate lineups (.43) than for fair lineups (.31). However, when presented sequentially the correct identification rate was higher in fair (.41) relative to intermediate (.24) lineups.

Although Carlson and colleagues' (2008) findings suggest a potential interaction between *E'* manipulations and lineup presentation format, the pattern did not replicate in subsequent research. Gronlund et al. (2009) used a combination of description-matching and appearance-

matching to construct 'medium' and 'fair' lineups that differed in *E'*. Large pools of potential fillers for the two lineup conditions were obtained by instructing research assistants to search an online mugshot database using different criteria. In the medium condition research assistants searched for fillers who matched the culprit on five descriptors, and then Gronlund selected five fillers from this pool who were neither very dissimilar nor very similar to the culprit's appearance, which produced *E'* values ranging from 2.33 to 3.15. In the fair condition different research assistants searched for fillers who matched the culprit on seven descriptors, and then Gronlund selected the fillers who best resembled the culprit, which produced *E'* values ranging from 3.75 to 4.51. This manipulation led to comparable innocent suspect misidentification rates and, contrary to the premise underlying the not-too-similar recommendation, the correct identification rate for fair lineups (.36) was higher than for medium lineups (.30), a trend that was consistent for simultaneous and sequential lineups.

Euclidean distance. Tredoux (2002) used a technique to manipulate similarity that is radically different than any of the previously described methods. Building on Valentine's (1991) euclidean multidimensional 'face space' framework, Tredoux used principal component analysis to systematically identify the euclidean distance (degree of similarity) between two faces. After establishing that this approach was comparable to perceptual similarity ratings, Tredoux (2002) conducted a lineup task comparing fillers who were very close in euclidean distance with fillers who were only moderately close. Tredoux did not specify separate accuracy rates for culprit-present and culprit-absent lineups; however, the similarity manipulation did not yield a significant difference in overall accuracy (moderate = .44; high = .40), and similarity did not interact with target-presence or lineup presentation.

Computer-generated faces. In contrast to each of previously-described experiments, which used photographs of real faces, Flowe and Ebbeson (2007) used a software program (FACES) to generate simulated faces that differed in similarity to a simulated target face. In the 'random similarity' condition, fillers were pseudo-randomly selected from a database of 1000 faces (eye colour was required to match; all other facial features were required to mismatch). In the 'matched similarity' condition, the lineup faces were matched on one facial feature. In two experiments, this manipulation yielded no significant differences in correct identifications. Consistent with the trend observed by Carlson et al. (2008), in Experiment 1 the correct identification rate for simultaneous presentation was higher for random similarity lineups (.52) than for matched similarity lineups (.45), but the correct identification rate for sequential presentation was lower for random similarity lineups (.33) than for matched similarity lineups (.37). The false identification rate for a 'look-a-like' in target-absent lineups was approximately 10% higher for random than matched similarity lineups, an effect that was uninfluenced by lineup presentation. Flowe and Ebbeson did not report the correct identification rates separately for the similarity conditions in Experiment 2, only noting the absence of a significant difference.

Summary of empirical findings. We reviewed 16 studies comparing lineups of moderate and high similarity to the target. Although concerns that highly similar fillers would make correct identifications too onerous appear valid for child witnesses, similarity's effect on adults' ability to correctly identify a culprit was much less clear. In one experiment, highly similar fillers were associated with a dramatic reduction in correct identifications (Wells et al., 1993). In the other studies, however, the effect of highly similar fillers on correct identifications has been equivocal. In some cases highly similar fillers were associated with a small decrease in correct identifications, whereas in other cases highly similar fillers were associated with a small

increase in correct identifications. When the literature shows mixed results, a meta-analysis can provide a clearer understanding of the effect.

Recent Meta-Analytic Findings on Suspect-Filler Similarity Effects

In a recent meta-analysis examining the effect of suspect-filler similarity on 6650 identification responses, increases in similarity generally corresponded with decreases in suspect identifications (Fitzgerald et al., 2013). In particular, increasing similarity facilitated a shift from suspect selections to filler selections, rather than to lineup rejections. In a comparison between low and moderate similarity lineups, the suspect-to-filler shift occurred in both culprit-present and culprit-absent lineups; however, in a comparison between moderate and high similarity lineups, the shift only occurred in culprit-absent lineups. That is, relative to moderate similarity fillers, high similarity fillers reduced misidentification of innocent suspects without impeding correct identifications of the culprit.

The finding that highly similar fillers reduced innocent suspect misidentifications is noteworthy. In our summary of the empirical literature, we merely suggested that the evidence showing a negative effect of highly similar fillers on correct identifications is not robust. In addition to supporting this assertion, the meta-analytic results went one step further, suggesting that highly similar fillers are actually beneficial. Innocent suspect misidentifications are dangerous because they confirm the investigator's suspicion of that suspect. Accordingly, a misidentified innocent suspect will continue to be investigated and could potentially be wrongfully convicted. By contrast, a misidentified filler will not be investigated because fillers are known innocents. Thus, although highly similar fillers did not increase the rate of correctly rejected culprit-absent lineups, the shift from innocent suspects to fillers would have the same exonerating effect in an applied setting.

Fitzgerald and colleagues (2013) concluded that high similarity lineups seemed to provide the best balance in terms of protecting innocent suspects and facilitating culprit identifications. However, they stopped short of a full endorsement of selecting the most similar fillers available, noting that their lineup categories were defined in relation to one another. Although suspect-filler similarity was greater in the high similarity category than in the moderate or low similarity categories, these were not objectively defined categories. Thus, Fitzgerald et al. suggested that researchers may not have constructed lineups with the degree of similarity that has been cautioned against.

The Present Research

In previous research, the ability to create high similarity lineups may have been limited by the availability of fillers who strongly resemble the suspect. To circumvent this issue, we manipulated suspect-filler similarity using morphing software. The software is capable of creating lineups with an extremely high degree of similarity, which is critical for identifying the point at which fillers resemble the suspect to such a degree that correct identifications become too difficult. We carefully pilot-tested the materials to ensure that morphing per se did not influence responding and that participants actually perceived greater similarity in the lineup manipulated to be higher in similarity. In the experiment, witnesses viewed a video containing a target person and were subsequently asked to attempt his identification from lineups containing (a) the target or an innocent suspect and (b) moderately high or very high similarity fillers. We predicted that relative to the moderately high similarity fillers, the very high similarity fillers would reduce identifications of both the target and the innocent suspect.

Method

Participants

In total, 271 undergraduate students were recruited. Approximately half of these students participated in the main experiment (n = 137; M_{age} = 20.70 years, SD = 3.76; 109 women) and half participated in pilot studies (n = 134) that were required to prepare the study materials.

Materials

Video event. The 6-minute silent video began with a man and woman preparing to eat breakfast at a restaurant. They ordered food and received beverages, but got into an argument and left the restaurant before their food arrived. Outside the restaurant, the argument resumed. Although the man (culprit) managed to sneak in a kiss, the woman ultimately pushed him away, got into a car, and fled the scene.

Lineups. The five faces in Figure 1 were altered using *Fantamorph* software to create fillers for four lineups that varied in target-presence and suspect-filler similarity (Figure 2). The original faces were selected because of their match to the culprit's appearance and their suitability for morphing.

Culprit present. The culprit-present lineups contained faces of the culprit and five fillers. A graphical representation of the morphing procedures used in the moderately high and very high similarity lineups is provided in Figure 3.

In the moderately high similarity lineup, the fillers in Figure 1 were morphed with the culprit to create a new face that was 40% culprit and 60% filler. Although this procedure only changed the appearance of the fillers slightly from their unmorphed photograph, morphing was

performed nonetheless to avoid having a very high similarity lineup that contained morphed fillers and a moderately high similarity lineup that contain unmorphed fillers.

A simple method of producing fillers who resemble the culprit more than the fillers in the moderately high similarity lineup would be to increase the degree of morph with the culprit. For example, we could have created faces that were 70% culprit and 30% filler. However, increasing the morph to this degree would produce fillers that are indistinguishable from both the culprit and each other. One author who was acquainted with the person acting as the culprit could not even identify the culprit from a lineup with such similar fillers.

To avoid this problem, fillers in the very high similarity lineup were morphed twice. First, we obtained similarity ratings to identify faces of high similarity to the culprit. Participants (n = 5) compared the culprit to 277 other faces and assigned a number from 0 (not at all similar) to 10 (highly similar). These data were used to select five faces of high similarity to the culprit, which were then morphed 50% with the faces in Figure 1. Each face was morphed with a different person of high similarity to the culprit to avoid having a particularly homogenous set of fillers, which could make the culprit stand out. Then, to ensure that the very high similarity fillers resembled the culprit more than the moderately high similarity fillers, the morphed faces were morphed again, with the culprit (40% culprit and 60% filler). This procedure produced fillers who were very similar to the culprit, but not so similar that the author who was acquainted with the culprit could not distinguish between the culprit and the fillers.

Although pilot testing (the relative judgement task, described below in the Pilot Tests section) indicated our procedure produced an effective difference in culprit-filler similarity between the moderately high and very high similarity conditions for three of the fillers, the difference in culprit-filler similarity between the moderately high and very high similarity

conditions was equivocal for Fillers C and E. This was addressed by adjusting the morph for Fillers C and E in the very high similarity condition to create faces that were 50% culprit and 50% filler instead of 40% culprit and 60% filler. Follow-up pilot tests indicated that this procedure created Fillers C and E who were more similar to the culprit in the very high similarity condition than in the moderately high similarity condition.

Culprit absent. Culprit-absent lineups comprised an innocent suspect and five fillers. Fillers were created by morphing the faces in Figure 1 with the innocent suspect's face. In previous studies, researchers have often used the same fillers in the culprit-present and culprit-absent lineups. That is, they simply replaced the culprit with an innocent suspect. By only changing one element of the lineup, this design has the advantage of high experimental control; however, this approach does not correspond with how the match-to-appearance procedure would be implemented by legal investigators constructing lineups for an innocent suspect because the appearance of the culprit would not be known (Clark & Tunnicliff, 2001). To simulate the lineup construction procedures that would occur in the field, we matched fillers in the culprit-absent lineups to the innocent suspect's appearance. When fillers are matched to an innocent suspect's appearance, the innocent suspect is hypothesized to resemble the culprit more than any of the fillers (cf., Navon, 1992; Wogalter, Marwitz, & Leonard, 1992). To create this effect in our lineups, the innocent suspect's face was morphed with the culprit's face (50%) and the filler faces were not morphed with the culprit's face.

The morphing procedure for fillers in the culprit-absent lineup was similar to that employed for fillers in the culprit-present lineup. In the moderately high similarity lineup, faces in Figure 1 were morphed to create a face that was 40% innocent suspect and 60% filler. In the very high similarity lineup, faces were morphed twice. First, each of the faces in Figure 1 was

morphed 50% with a unique face that was judged to bear a strong resemblance with the innocent suspect in pilot testing (n = 4; same rating procedure as was used with the culprit). The resulting faces were then morphed with the innocent suspect (40% innocent suspect and 60% filler).

Pilot testing (the relative judgement task) indicated this procedure did not produce an effective difference in suspect-filler similarity between three pairs of moderately high and very high similarity fillers (Fillers B, C, and E). This was addressed by morphing these three very high similarity fillers with the innocent suspect 50% instead of 40%. For each of the three pairs, follow-up pilot tests indicated this procedure created very high similarity fillers who resembled the innocent suspect more than did the moderately high similarity fillers.

Pilot Tests

The effectiveness of the morphing software in producing a manipulation of similarity was evaluated in a series of pilot tests.

Relative judgement task. In the first set of pilot studies, participants made relative judgements about which of two lineup members (a moderately high similarity filler vs. a very high similarity filler) was more similar to the suspect. In the first pilot study, judges (n = 12) completed multiple trials in which three faces were presented in a row. The suspect (i.e., the culprit or the innocent suspect) was always in the middle position. Fillers were always positioned on the left and right sides. The number “1” was always displayed over the face on the left side and the number “0” was always displayed over the face on the right side. Judges were instructed to determine whether Person 0 or Person 1 looked more similar to the person in the middle. For example, the culprit would be in the middle, Filler A from the moderately high similarity culprit-present lineup would be on the left, and Filler A from the very high similarity culprit-present lineup would be on the right. If the very high similarity filler resembles the culprit more than

does the moderately high similarity filler, participants should choose the very high similarity filler more frequently.

For the similarity manipulation to be acceptable, we set an arbitrary criterion stating that the very high similarity filler needed to garner at least twice as many choices as the moderately high similarity filler. This criterion (i.e., 8/12) was met in 5 of the 10 comparisons. In the other five comparisons, the number of choices for the moderately high and very high similarity fillers was approximately evenly split. This was addressed by making a slight increase in the extent to which very high similarity fillers were morphed with the suspect and then conducting a follow-up pilot study with new participants (n = 12) to confirm the new faces met the previously mentioned criterion. The modifications to these fillers are detailed in the Materials section.

Find the nonmorph task. In the culprit-present lineups, the photograph of the culprit is the only nonmorphed photograph. Thus, witnesses could potentially choose the culprit not because they remember him from the video, but rather because they can tell the other faces have been altered in some way and that the culprit's photograph is the only one that has not been digitally altered. To assuage such concerns, pilot studies were conducted with the culprit-present lineups to ensure that the culprit's face could not be distinguished from the fillers in the absence of recognition.

In the pilot studies, participants (n = 84) who did not see the crime video were informed they would view a simultaneous lineup containing five digitally manipulated faces and one unaltered face (location of culprit was counterbalanced). Their task was to choose the face that had not been manipulated. For the 43 participants who viewed the moderately high similarity lineup, the proportion (P) who chose the culprit (P = .16, SE = .06) did not differ from chance

(.17). For the 41 participants who viewed the very high similarity lineup, the proportion who chose the culprit (P = .20, *SE* = .06) also did not differ from chance.

Suspect-filler similarity ratings. The similarity between the suspects and the morphed fillers was further assessed through subjective similarity ratings. A new set of judges (*n* = 17) viewed suspect-filler face pairs and assigned a number from 0 (not at all similar) to 10 (highly similar). These judges also rated the similarity between the culprit and the innocent suspect (*M* = 6.53, *SE* = 0.61). Paired-samples *t*-tests were used to assess the difference between ratings for a filler in the moderately high similarity condition and for the corresponding filler in the very high similarity condition. Table 2 shows reliable differences in the predicted direction for Fillers A, B, C, and D, but no reliable differences for Filler E. Additional paired-samples *t*-tests were used to contrast average ratings for the five fillers in the moderately high and very high similarity conditions. For culprit-present and culprit-absent conditions, the moderately high similarity fillers were rated to be significantly less similar than the very high similarity fillers.

Procedure

To prevent knowledge of an upcoming memory test, the study was advertised as an investigation of media influences on gender roles. Upon arrival, the experimenter informed participants that they would watch a video and then answer questions about gender roles. Participants then viewed the video described in the Materials section, which was presented on a 21-inch computer screen. After the viewing, the experimenter asked participants to report what happened in the video. No follow-up questions were asked and no feedback was given. Participants were subsequently informed that the experiment was actually about eyewitness identification, not gender roles. Participants were instructed to imagine that the man from the video committed a crime and that they were the only person in a position to identify him.

Participants were then asked if they were willing to attempt a lineup identification. All participants consented.

To facilitate double-blind lineup administration, the lineup task was completed on a computer. The experimenters who opened the computer program did not know whether the culprit was in the lineup, nor did they know which fillers the lineup contained. Before participants started the computer task, the experimenter verbally instructed them that the culprit may or may not be present. He further instructed participants that they could reject the lineup if they did not think the culprit was present.

Lineup members were presented simultaneously in a 2 × 3 array. Each face was associated with a number (1-6). The spatial location of the suspect's photograph was counterbalanced. The computer program instructed participants that if the man from the video was present, they were to press the number associated with his image. If absent, participants were instructed to press '0'. After the identification task, participants provided a confidence assessment about their identification decision using a 6-point scale, ranging from 1 ("not all at confident") to 6 ("highly confident"). In an exit interview, participants were asked to report their demographic information, previous experience participating in eyewitness experiments, and whether they knew that the experiment involved an identification test before viewing the video.

Results

The exit interview revealed seven participants who claimed awareness that the experiment involved an identification test before watching the video. These seven participants were omitted from all data analyses, which had no impact on whether any of the differences were significant.

Effects of Similarity on Identification Responses

The significance of associations between categorical variables was assessed with *z*-tests for the difference between two proportions. For effect size measures, odds ratios (OR) with 95% confidence intervals are reported with all *z*-tests. An odds ratio of 1.00 indicates perfect unity between two groups in the odds of an identification response between conditions (Bland & Altman, 2000). For the analyses that follow, an odds ratio above 1.00 indicates a greater likelihood of a response in the moderately high similarity condition than in the very high similarity condition.

On culprit-present lineups, similarity was associated with suspect and filler choices (Table 3). A greater proportion of culprit identifications were made when similarity was moderately high relative to very high, $z = 2.48$, $p = .01$, OR = 4.04, 95% CI = 1.24, 13.23. Correspondingly, a smaller proportion of filler selections were made in the moderately high similarity condition relative to the very high similarity condition, $z = 2.77$, $p = .01$, OR = 0.25, 95% CI = 0.09, 0.71. Similarity had no effect on incorrect lineup rejections, $z = 0.49$, $p = .62$, OR = 1.34, 95% CI = 0.43, 4.20.

On culprit-absent lineups, the innocent suspect was misidentified more frequently from the moderately high similarity lineup than from the very high similarity lineup (Table 3); however, the difference was small and nonsignificant, $z = 0.62$, $p = .54$, OR = 1.41, 95% CI = 0.48, 4.19. The only significant effect in the culprit-absent condition was for filler selection rates, which were lower for moderately high relative to very high similarity lineups, $z = 2.54$, $p = .01$, OR = 0.22, 95% CI = 0.06, 0.78. The correct rejection rate was higher for the moderately high similarity lineup than for the very high similarity lineup; however, the difference did not reach significance, $z = 1.59$, $p = .11$, OR = 2.19, 95% CI = 0.83, 5.83.

To explore similarity effects on only those who chose a lineup member (choosers), the analyses were repeated after excluding those who rejected the lineup. Among choosers, similarity had comparable effects on culprit-present and culprit-absent lineups (Figure 4). In particular, both the culprit ($z = 2.99$, $p = .003$, OR = 5.91, 95% CI = 1.62, 21.54) and the innocent suspect ($z = 2.01$, $p = .04$, OR = 4.06, 95% CI = 0.95, 17.43) were more likely to be chosen from moderately high relative to very high similarity lineups. Thus, regardless of whether the culprit was present or absent, increasing similarity for choosers resulted in a shift from identification of the suspect to identification of a filler.

Diagnosticity

Diagnosticity ratios, which are the ratio of an identification response probability in culprit-present and culprit-absent lineups, can be calculated to assess the extent to which a response is indicative of the suspect's guilt or innocence (Wells & Lindsay, 1980). Diagnosticity ratios for suspect identifications are typically used to provide information about guilt, whereas diagnosticity ratios for filler identifications and lineup rejections are typically used to provide information about innocence (Wells & Olson, 2002).

Suspect selections. Diagnosticity ratios for suspect selections were calculated by dividing the culprit selection rate by the innocent suspect selection rate. A ratio above unity (i.e., 1.00) would suggest that suspect identifications from a given lineup are indicative of guilt. A larger diagnosticity ratio was found for the moderately high similarity lineups (1.45, 95% CI = 0.75, 2.75) than for very high similarity lineups (0.69, 95% CI = 0.25, 1.87). Thus, a suspect identified from a moderately high similarity lineup was indicative of guilt, whereas a suspect identified from a very high similarity lineup was not indicative of guilt.

Filler selections and lineup rejections. Diagnosticity ratios for filler selections and lineup rejections were calculated by dividing the response rate for the culprit-absent lineup by the response rate for the culprit-present lineup. A ratio above unity (i.e., 1.00) would suggest that filler and rejection responses from a given lineup are indicative of innocence. For filler selections, the ratios in both similarity conditions were less than 1.00 (moderately high = 0.43, 95% CI = 0.14, 1.26; very high = 0.62, 95% CI = 0.37, 1.04). For lineup rejections, the ratios in both similarity conditions were greater than 1.00 (moderately high = 2.05, 95% CI = 1.09, 3.83; very high = 1.69, 95% CI = 0.78, 3.69). Thus, rejections were indicative of innocence, but filler selections were not.

Confidence

Influences on post-identification confidence were assessed with two 2 (lineup similarity) × 2 (identification accuracy) analyses of variance (ANOVA): one for culprit-present lineups and one for culprit-absent lineups (Figure 5). The ANOVA for culprit-present lineups revealed a main effect of similarity, $F(1,59) = 8.03$, $p = .006$, $d = 0.89$, 95% CI = 0.60, 1.17, indicating greater confidence for the moderately high similarity lineups relative to the very high similarity lineups. There was no main effect of accuracy and no interaction. The ANOVA for culprit-absent lineups revealed no main effects and no interaction.

Consistent with previous research (Sporer, Penrod, Read, & Cutler, 1995), a stronger confidence-accuracy correlation was observed for choosers ($r = .29$, $p = .007$) relative to nonchoosers ($r = -.01$, $p = .93$); however, the difference between the correlations (.30; 95% CI = -.06, .64) did not reach significance, $z = 1.65$, $p = .10$. When calculated only for moderately high similarity lineups, the confidence-accuracy relation was much stronger for choosers ($r = .38$, $p = .02$) than for nonchoosers ($r = -.24$, $p = .21$) and the difference between correlations (.62; 95% CI

= .20, .97) was significant, $z = 2.46$, $p = .01$. Conversely, when similarity was very high, the relation was stronger for nonchoosers ($r = .30$, $p = .20$) than for choosers ($r = .00$, $p = 1.00$), and the difference between correlations (.30; 95% CI = -.25, .76) was nonsignificant, $z = 1.08$, $p = .28$.

Discussion

Our primary objective was to establish the upper bounds of suspect-filler similarity. Consistent with our prediction, participants were much more adept at identifying the culprit from the moderately high similarity lineup than from the very high similarity lineup. The correct identification rate for the very high similarity lineup was quite low, which suggests the morphing software was successful at establishing the degree of similarity required to impede correct identifications. The decrease in correct identifications corresponded with an increase in filler identifications, suggesting the very high similarity fillers drew choices away from the culprit. These data suggest suspect-filler similarity can be 'too high', at least with the use of morphing software.

Only partial support was found for the hypotheses concerning similarity's effect on culprit-absent lineups. Although the increase in suspect-filler similarity led to a concomitant increase in filler selections, it yielded only a small and nonsignificant reduction in the overall rate of innocent suspect misidentifications. However, these rates were skewed by different choosing rates across similarity conditions. When only choosers were considered, both the innocent suspect and the culprit were significantly more likely to be identified from the moderately high similarity lineup than from the very high similarity lineup. Thus, costs and benefits were associated with very high similarity fillers.

The absence of an effect of similarity on rejection rates was one of the most consistent findings in a meta-analysis of suspect-filler similarity, so the difference in correct rejection rates between the moderately high (58%) and very high (38%) similarity lineups in the present research was not anticipated (Fitzgerald et al., 2013). Fitzgerald et al. used Clark's (2003) WITNESS model to interpret the null effect of similarity on choosing/rejection rates. According to WITNESS, lineup choices may occur because one lineup member is a strong match with the witness's memory of the culprit or because one lineup member is a much better match than any of the other lineup members. Fitzgerald et al. noted that as suspect-filler similarity increases, two effects on choosing were possible: (1) choosing could increase because of the higher likelihood that one of the lineup members would exceed a threshold for choosing, or (2) choosing could decrease because of the reduced difference between best and next-best matches.

Fitzgerald et al. (2013) suggested that if these two competing effects are of equal strength, similarity would have no effect on choosing. However, in the present research one effect may have been stronger than the other. In particular, morphing the very high similarity fillers with the innocent suspect may have increased the likelihood that one of the lineup members exceeded the decision criterion for making a positive selection. Although the increase in suspect-filler similarity should also have decreased the difference between the best and next-best matches, this might not have had any effect because the resemblance between the suspect and the fillers was relatively high in both the moderately high and very high similarity conditions. In other words, the difference between the best and next-best matches may not have been sufficiently large in either of the similarity conditions to warrant a positive identification, thus forcing witnesses to base their decision on whether the recognition experience elicited by one of the lineup members exceeded their decision threshold. Given the increased chance of a

lineup member in the very high similarity lineup exceeding this threshold, this interpretation is consistent with the increased rate of choosing in the very high similarity condition.

Diagnosticity

Suspect selections. The diagnosticity ratios indicated that suspect identifications from the moderately high similarity lineup were indicative of guilt, but suspect identifications from the very high similarity lineup were not. This is interesting because diagnosticity typically increases as fillers become more similar to the suspect (Fitzgerald et al., 2013). We suspect that the 'high similarity' fillers in previous research were not as similar as the fillers we created through morphing software, which could explain the atypical effect of similarity on diagnosticity. Consistent with this interpretation, similarity had a minimal effect on innocent suspect misidentifications and the lower diagnosticity ratio for the very high similarity lineups was primarily reflective of a drop in correct identifications. In other words, reducing similarity to only moderately high made it easier to identify the target.

Reducing the similarity between fillers and the innocent suspect might be expected to cause an increase in innocent suspect selections, so it is noteworthy that this did not happen. For this issue, two points are worthy of note. First, the match between the suspect's photograph and the memory of the culprit should typically be greater for the culprit than for an innocent suspect. Therefore, even if suspect-filler similarity were identical across target-present and target-absent lineups, the difference between the best match and the next best match would be greater for target-present than target-absent lineups. Second, the similarity ratings in Table 2 suggest that suspect-filler similarity was not identical across target-present and target-absent lineups. Although we used comparable procedures for constructing target-present and target-absent lineups, suspect-filler similarity for the moderately high similarity condition was rated higher for

the target-absent lineup (M = 5.39) in comparison to the target-present lineup (M = 4.39). The combination of these two factors can be expected to have created a larger difference between the best and the next-best match in the target-present condition than in the target-absent condition, which would explain the asymmetrical effect of similarity on suspect identifications in moderately high similarity lineups.

It should also be noted that diagnosticity ratios for suspect identifications in both similarity conditions (0.69-1.45) were substantially smaller than the average diagnosticity ratio of 10.00 for high similarity lineups in previous research (Fitzgerald et al., 2013). The size of a diagnosticity ratio is heavily influenced by the innocent suspect misidentification rate. If a lineup or procedure leads to low innocent suspect misidentification rates, larger diagnosticity ratios can be expected (Wixted & Mickes, 2012). Given the high innocent suspect misidentification rates that we observed, the small diagnosticity ratios should come as no surprise.

Filler selections and lineup rejections. Diagnosticity ratios for filler and rejection responses were relatively unaffected by lineup similarity. In both conditions, rejections provided information about the suspect's innocence, whereas filler selections were actually indicative of the suspect's guilt. These results can be explained as a product of the lineup construction procedures.

The diagnosticity values for rejections were above unity, indicating that a rejection was more likely to occur for a culprit-absent lineup than for a culprit-present lineup. Fillers in the culprit-absent condition were matched to the innocent suspect's appearance, as opposed to the common procedure of using the same fillers in culprit-present and culprit-absent lineups. This procedure should have the effect of making fillers in the culprit-present lineup match the representation of the culprit in memory better than the fillers in the culprit-absent lineup. As a

consequence, eyewitnesses will be more likely to reject the culprit-absent lineups than the culprit-present lineups.

The diagnostic values for filler selections were below unity, indicating that a filler selection is more likely to occur if the culprit is present than if the culprit is absent. Given that culprit-present fillers were matched to the culprit's appearance, they can be expected to match the memorial representation of the culprit to a greater extent than fillers who were matched to the innocent suspect's appearance. This finding is consistent with previous research showing that filler selections are indicative of the suspect's innocence for the match-to-description procedure, but not for the match-to-appearance procedure (Clark & Wells, 2008).

Confidence

According to models of information accumulation (Van Zandt, 2000; Vickers, 1979), confidence is a product of the difference between the recognition experience elicited by the chosen and non-chosen options. In the very high similarity condition, the extreme homogeneity of the lineup members could be expected to produce only a small difference in the feeling of familiarity between the culprit and the fillers, which would lead to uncertainty in the witnesses. In the moderately high similarity condition, the culprit could be expected to appear notably more familiar than the fillers and lead to greater certainty than in the very high similarity condition. Consistent with the model of information accumulation, participants in the culprit-present condition were more confident in their identification responses for moderately high similarity lineups than for very high similarity lineups. Similarity had no effect on confidence in the culprit-absent condition, which could indicate that the difference in familiarity between the innocent suspect and the fillers was not sufficiently greater in the moderately high similarity condition than in the very high similarity condition to affect confidence.

Potential Applications

Surveys consistently show that the majority of investigators in the field (81-83%) select fillers by matching to the suspect's appearance (Police Executive Research Forum, 2013; Wogalter, Malpass, & McQuiston, 2004). In the more recent survey, respondents indicated using the match-to-appearance procedure in different ways. Of the 81% of agencies that reported using a match-to-appearance procedure, 50% indicated using fillers who were matched on "the general characteristics of the suspect" (p. 59) and 31% indicated using "fillers who look as much like the suspect as possible" (p. 59). Presumably, police agencies would be keen to learn which of these two methods is best supported by empirical research.

At this point, arguments in favour of either strategy could be made. Fitzgerald and colleagues' (2013) meta-analysis indicated that relative to moderate similarity lineups, high similarity lineups reduce innocent suspect misidentifications without reducing correct identifications. This finding could be interpreted as support for the construction of lineups with the most similar fillers available. However, the present findings suggest a strategy of matching as much as possible could be problematic. We found that if fillers resemble the suspect too much, the correct identification rate is substantially reduced and the overall innocent suspect misidentification rate is only slightly reduced. Of course, whether investigators in the field could create lineups akin to our very high similarity lineups is debatable. Previous research with unmorphed photographs has demonstrated that highly similar fillers can reduce correct identifications (Sauer et al., 2008; Wells et al., 1993); however, our literature review suggests the results of these experiments seem to be the exception rather than the rule.

Relying on experimental research to decide whether fillers should be matched on a general description or as closely as possible to the suspect is further complicated by the fact that

researchers rarely construct lineups using procedures comparable to standard police practice. The morphing methodology is undeniably quite different from the typical lineup construction procedures utilized in applied settings, and it could have led to fillers who were more similar than the most similar fillers that would be available to police. However, researchers using more traditional lineup construction strategies often report selecting photographs from relatively small face databases. Police typically have access to hundreds of thousands of mugshots or driver's license photographs (Police Executive Research Forum, 2013), so the most similar fillers available in researchers' face databases may not be as similar as those available to police investigators. Consistent with this assertion, after finding no difference in correct identifications in their meta-analytic comparison between moderate and high similarity lineups, Fitzgerald et al. (2013) questioned whether the lineups operationally defined as 'high similarity' contained the degree of similarity that had been cautioned against.

The morphing methodology was important for establishing a boundary condition for matching fillers to suspects. However, a policy recommendation on whether police should or should not be selecting the most similar fillers available needs to be grounded in research using the same resources that are available to those constructing lineups in the field. Although we are not aware of any empirical research using drivers' license photographs, mugshot databases have been used in a small number of empirical studies. For instance, Gronlund et al. (2009) selected fillers from the Florida Supervised Offenders database, which contains more than 100,000 mugshots. A detailed description of Gronlund and colleague's lineup construction procedures is provided in the Introduction, so rather than repeating it here we simply note that quite rigorous procedures were used to create 'fair' lineups that comprised fillers who resembled the suspect and correct identifications were still relatively commonplace. This suggests a strategy of

selecting the most similar fillers may be tenable, but additional research on lineup construction procedures using police resources is necessary before a firm policy recommendation can be advocated.

Limitations and Future Directions

Although the morphing methodology has numerous advantages, it also has limitations. First, not all face pairs can be easily morphed into a new, natural-looking face. To obscure the fact that our images had been morphed, we had to select images that were already somewhat homogenous (e.g., short hair). Because of this restriction, the morphing methodology may not be a viable option for all investigations of lineup member similarity.

Second, practical issues arise when constructing culprit-absent lineups with morphed faces. Previous research has indicated that matching fillers to the suspect's appearance results in a lineup containing a suspect who would resemble the culprit more than would any of the fillers (Navon, 1992; Wogalter et al., 1992). In our investigation, we simulated this effect by morphing the innocent suspect with the culprit; however, the rather high innocent suspect misidentification rates suggest this approach created a lineup containing a greater degree of bias than that which would have been present in lineups constructed through less artificial means. For this reason, we recommend caution when interpreting the present results in the culprit-absent condition and encourage future research to experimentally examine the impact of having such a highly similar innocent suspect.

Third, manipulating similarity with morphing software will always raise concerns about ecological validity. Although our pilot work suggests nonmorphed faces can be difficult to detect from among a set of morphed faces, the use of computer software to manipulate similarity is an inherently artificial method of imitating differences in lineup composition that might arise from

different filler selection strategies. Fillers in the very high similarity condition seem to have made culprit identifications extremely difficult, but whether police have the resources to create lineups with such high similarity is uncertain. Moving forward, researchers will need to construct lineups using police databases to ascertain whether natural filler selection procedures are capable of creating the level of similarity that was present in our lineups.

Summary and Conclusions

We used morphing software to create lineups with varying degrees of suspect-filler similarity. We assigned the labels "moderately high" and "very high" to indicate our position that the lineups in both similarity conditions contained highly similar fillers. Our findings suggest lineups that are comparable to our very high similarity lineups have the potential to hinder correct identifications. However, future research is needed to determine whether such lineups would occur as a result of conventional filler selection procedures. We suspect the most similar fillers available will be appropriate in most instances; however, this would almost certainly depend on the size of the filler database. Most driver's license databases should be quite large, so police using such databases may need to exercise caution when selecting the most similar fillers. Given recent findings associated with criminal face biases (Flowe & Humphries, 2011), we are also interested in whether suspect identification rates are influenced by placing a suspect (who is likely to have a criminal past) in a lineup with fillers chosen from a driver's licence photograph database (who are less likely to have a criminal past). Accordingly, we encourage future researchers to examine the influence of filler database characteristics on the composition of lineups and the identification responses of eyewitnesses.

References

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *BMJ, 320*, 1468. DOI: 10.1136/bmj.320.7247.1468

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15*, 77-96. DOI: 10.1348/135532509X414765

Brewer, N., & Wells, G. L. (2006). The confidence–accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11-30. DOI: 10.1037/1076-898X.12.1.11

Brigham, J. C., & Brandt, C. C. (1992). Measuring lineup fairness: Mock witness responses versus direct evaluations of lineups. *Law and Human Behavior, 16,* 475-489.

Brigham, J. C., Ready, D. J., & Spier, S. A. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology, 11,* 149-163

Burton, A. M., White, D., & McNeill, A. (2010). 'The Glasgow Face Matching Test'. *Behavior Research Methods*, 42, 286-291. DOI: 10.3758/BRM.42.1.286

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 14,* 118-128. DOI:10.1037/1076-898X.14.2.118

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17,* 629-654. DOI:10.1002.acp.891

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16,* 22-42. DOI: 10.3758/PBR.16.1.22

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior, 32,* 187-218. DOI:10.1007/s10979-006-9082-4

Clark, S. E., & Tunnicliff, J. L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior, 25,* 199-216. DOI:10.1023/A:1010753809988

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior, 32,* 406-422. DOI 10.1007/s10979-007-9115-7

Darling, S., Valentine, T., & Memon, A. (2008). Selection of lineup foils in operational contexts. *Applied Cognitive Psychology, 22,* 159-169. DOI: 10.1002/acp.1366

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups: partial remembering. *Journal of Police Science and Administration, 18,* 287-293.

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19,* 151-164. DOI: 10.1037/a0030618

Fitzgerald, R. J., Whiting, B. F., Therrien, N. M., & Price, H. L. (2014). Lineup member similarity effects on children's eyewitness identification. *Applied Cognitive Psychology*. Advance online publication. DOI: 10.1002/acp.3012

Flowe, H. D., & Ebbesen, E. B. (2007). The effect of lineup member similarity on recognition accuracy in simultaneous and sequential lineups. *Law and Human Behavior, 31,* 33-52. DOI: 10.1007/s10979-006-9045-9

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology, 25,* 265-273. DOI: 10.1002/acp.1673

Glass, A. L., Holyoak, K. J., & Santa, J. L. (1979). *Cognition*. Reading, MA: Addison-Wesley.

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15,* 140-152. DOI: 10.1037/a0015082

Guttman, L., & Schlesinger, E. M. (1967). Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement, 27,* 569-580.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1304-1316.

Lindsay, R. C. L., Mansour, J. K., Beaudry, J. L., Leach, A. M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal and Criminological Psychology, 14,* 13-24. DOI: 10.1348/135532508X382708

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.

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4,* 303-313.

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

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior, 5,* 299-309.

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup

fairness. In R. Lindsay, D. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology, Volume 2: Memory for people* (155-178). Mahwah, NJ: Erlbaum.

McQuiston-Surrett, D., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: A review of methods, data, and theory. *Psychology, Public Policy, and Law, 12,* 137-169. DOI: 10.1037/1076-8971.12.2.137

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behavior, 16,* 575-593. DOI:10.1007/BF01044624

Police Executive Research Forum. (2013). *A National Survey of Eyewitness Identification Processes in Law Enforcement Agencies*. Washington, DC: Author. Retrieved from http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf

Sauer, J. D., Brewer, N., Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137,* 528-547. DOI: 10.1037/a0012712

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature, 470,* 437. DOI:10.1038/470437a

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25,* 259–274.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness studies. *Psychological Bulletin, 118,* 315-327.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99-139. DOI: 10.1037/a0021650

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22,* 217-237.

Tredoux, C. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. *Journal of Experimental Psychology: Applied, 8,* 180-193. DOI: 10.1037/1076-898X.8.3.180

Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions? *Law and Human Behavior, 24,* 231-258.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43,* 161-204. DOI: 10.1080/14640749108400966

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 26,* 582-600. DOI: 10.1037//0278-7393.26.3.582

Vickers, D. (1979). Decision processes in visual perception. New York: Academic Press.

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 3,* 776-784.

Wells, G. L., & Luus, C. A. E. (1990). The diagnosticity of a lineup should not be confused with the diagnostic value of non-lineup evidence. *Journal of Applied Psychology, 75,* 511-516.

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 3,* 155-167. DOI:10.1037/1076-898X.8.3.155

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.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7,* 275-278. DOI:10.1177/1745691612442906

Wogalter, M. S., Malpass, R. S., & McQuiston, D. E. (2004). A national survey of U. S. police on preparation and conduct of identification lineups. *Psychology, Crime & Law, 10,* 69-82. DOI: 10.1080/10683160410001641873

Wogalter, M. S., Marwitz, D. B., & Leonard, D. C. (1992). Suggestiveness in photospread lineups: Similarity induces distinctiveness. *Applied Cognitive Psychology, 6,* 443-453. DOI:10.1002/acp.2350060508

Table 1

Correct and false identification (ID) rates for Description-Matched (DM) and Appearance-Matched (AM) lineups

Experiment	*n*	Correct ID Rate		*z*	*p*	Odds Ratio & 95% CIs		
		DM	AM			OR	Lower	Upper
Wells et al. (1993)	84	.67 (.07)	.21 (.06)	4.73	.001	7.33	2.76	19.48
Lindsay et al. (1994)	58	.79 (.08)	.66 (.09)	1.10	.271	2.01	0.62	6.57
Juslin et al. (1996)	192	.52 (.05)	.44 (.05)	1.10	.268	1.40	0.79	2.47
Tunnicliff & Clark (2000) Exp. 1	64	.53 (.09)	.53 (.09)	0.00	1.000	1.00	0.38	2.67
Tunnicliff & Clark (2000) Exp. 2[1]	48	.31 (.07)	.33 (.07)	-	-	-	-	-
Darling et al. (2008)	100	.45 (.07)	.49 (.07)	0.40	.692	0.86	0.39	1.90
		False ID Rate						
		DM	AM					
Wells et al. (1993)	84	.12 (.05)	.12 (.05)	0.00	1.000	1.00	0.27	3.75
Lindsay et al. (1994)	137	.15 (.04)	.04 (.02)	2.11	.035	3.92	1.03	14.93
Juslin et al. (1996)	64	.09 (.05)	.09 (.05)	0.00	1.000	1.00	0.19	5.37
Tunnicliff & Clark (2000) Exp. 1	64	.13 (.06)	.03 (.03)	1.40	.161	4.43	0.47	42.02
Tunnicliff & Clark (2000) Exp. 2[1]	48	.08 (.04)	.19 (.06)	-	-	-	-	-
Darling et al. (2008)	100	.05 (.03)	.04 (.02)	0.29	.769	1.34	0.18	9.92

Note: Standard errors for ID rates are in parentheses.

[1] In Tunnicliff and Clark's second experiment, a within-subjects design was employed. This represents a violation of the independence assumption for the two-proportions *z* test. Thus, *z* and *p* were not computed.

Table 2

Mean (SE) ratings of suspect-filler similarity

		Similarity Condition				Cohen's *d* and 95% CIs		
Suspect	Filler	Moderately High	Very High	*t*	*p*	*d*	Lower Limit	Upper Limit
Culprit	A	5.37 (0.56)	7.63 (0.47)	3.65	.002	0.91	0.25	1.57
	B	3.71 (0.45)	6.00 (0.56)	3.43	.003	0.82	0.15	1.48
	C	4.88 (0.54)	7.50 (0.43)	6.15	.001	1.58	0.88	2.26
	D	3.56 (0.57)	6.31 (0.58)	3.32	.005	0.83	0.08	1.57
	E	4.47 (0.59)	5.00 (0.53)	1.28	.217	0.32	-0.04	0.67
	Average	4.39 (0.36)	6.50 (0.40)	5.63	.001	1.41	0.72	2.09
Innocent	A	5.31 (0.78)	8.06 (0.44)	3.51	.003	0.90	0.24	1.55
	B	6.06 (0.63)	7.81 (0.53)	2.43	.028	0.60	-0.02	1.22
	C	4.94 (0.60)	7.18 (0.55)	3.53	.003	0.86	0.25	1.41
	D	5.12 (0.55)	6.88 (0.49)	3.85	.001	0.93	0.42	1.45
	E	6.25 (0.54)	6.06 (0.57)	0.31	.764	-0.08	-0.44	0.60
	Average	5.39 (0.47)	7.08 (0.42)	5.64	.001	1.46	0.83	2.08

Note. Scale ranged from 0 (Not at all similar) to 10 (highly similar).

Table 3

Identification response probabilities (P) and standard errors (SE) as a function of culprit-presence and suspect-filler similarity

Culprit	**Similarity**	***n***	**Identification Response**					
			Suspect		**Filler**		**Reject**	
			P	*SE*	P	*SE*	P	*SE*
Present	Moderately High	32	.44	.09	.28	.08	.28	.08
	Very High	31	.16	.07	.61	.09	.23	.08
Absent	Moderately High	33	.30	.08	.12	.06	.58	.09
	Very High	34	.24	.07	.38	.08	.38	.08

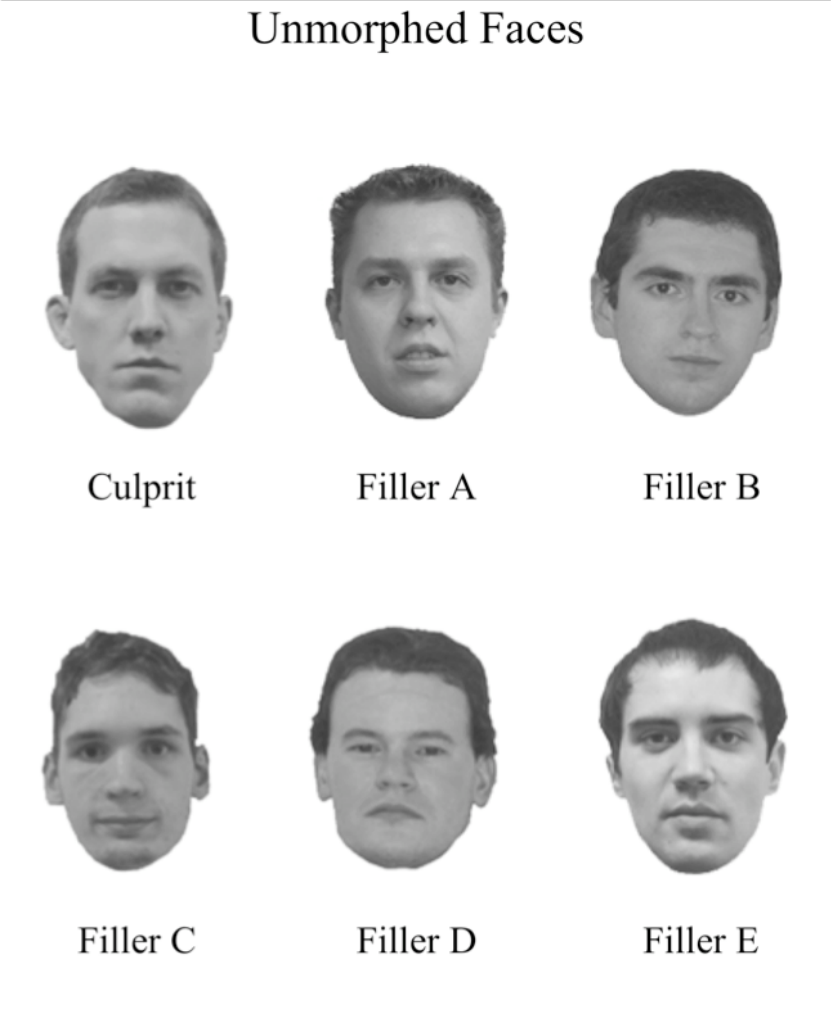

Figure 1. Faces of the culprit and the fillers (prior to the morphing procedure).

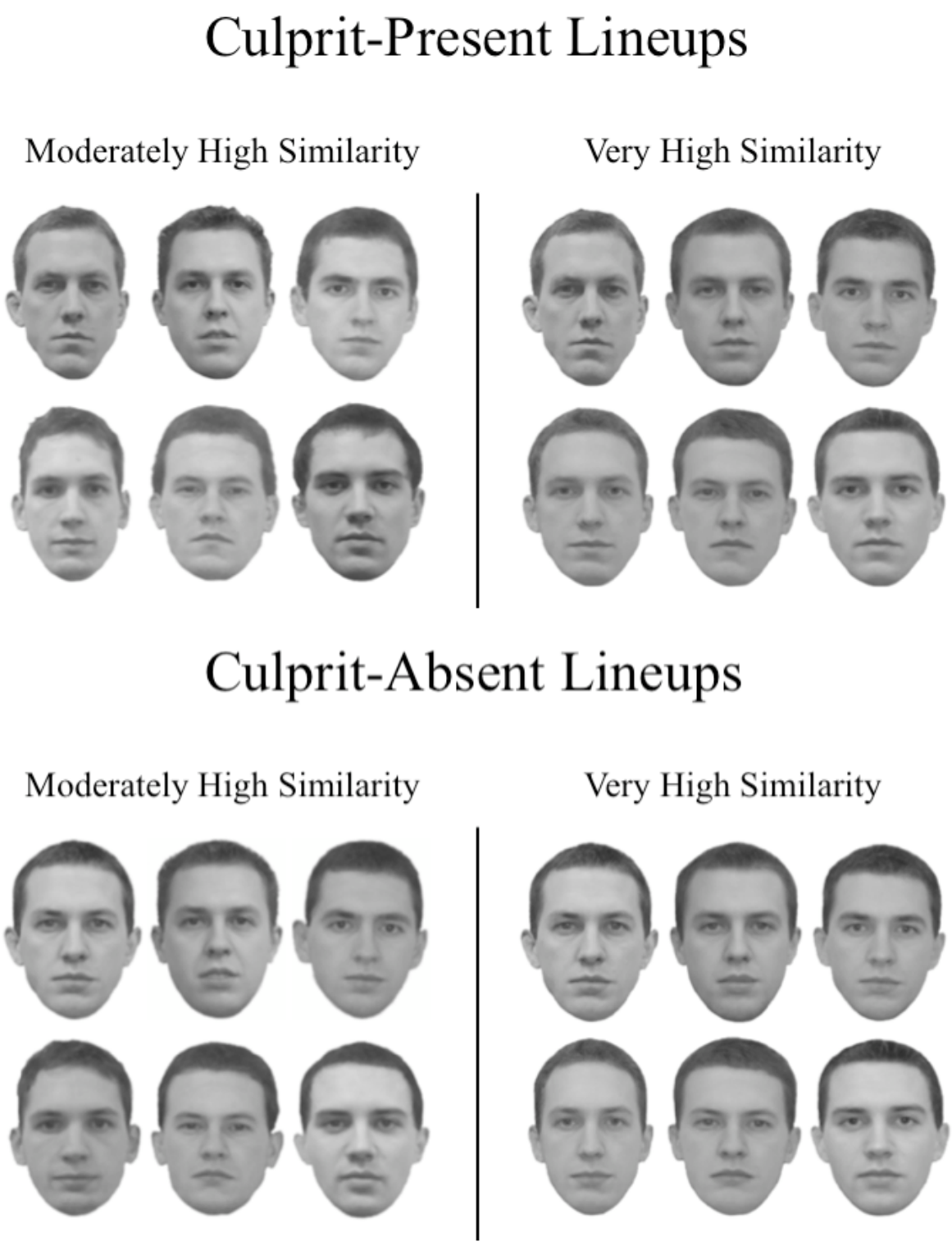

Figure 2. Lineups for moderately high and very high similarity fillers. In these lineups, the suspect is always positioned in the top-left corner. In the experiment, suspect position was counterbalanced.

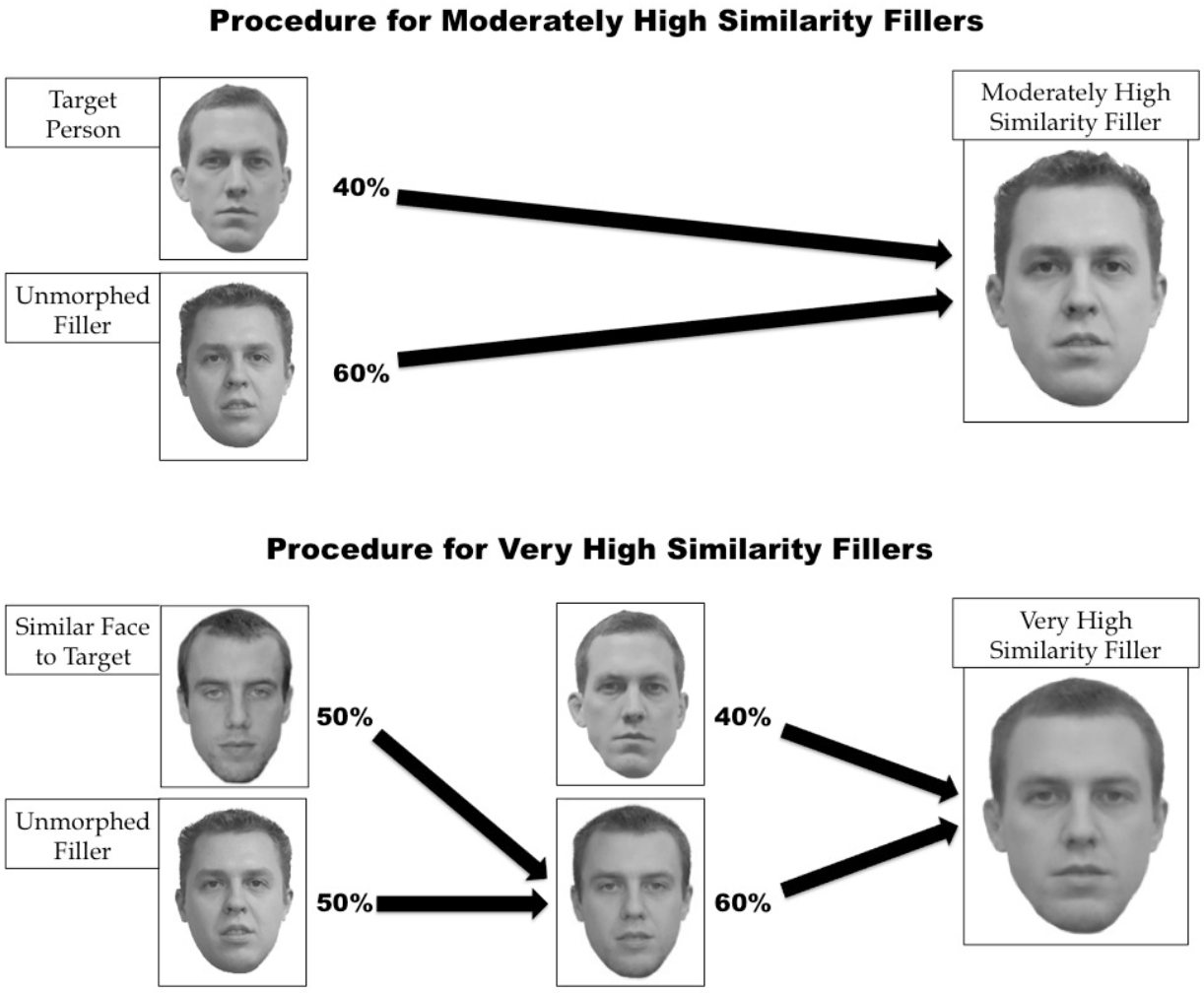

Figure 3. Standard morphing procedures used to create fillers in moderately high and very high similarity conditions. The target and filler images were photographed locally. The 'Similar Face to Target' was obtained from the Glasglow Unfamiliar Face Database (Burton, White, & McNeill, 2010).

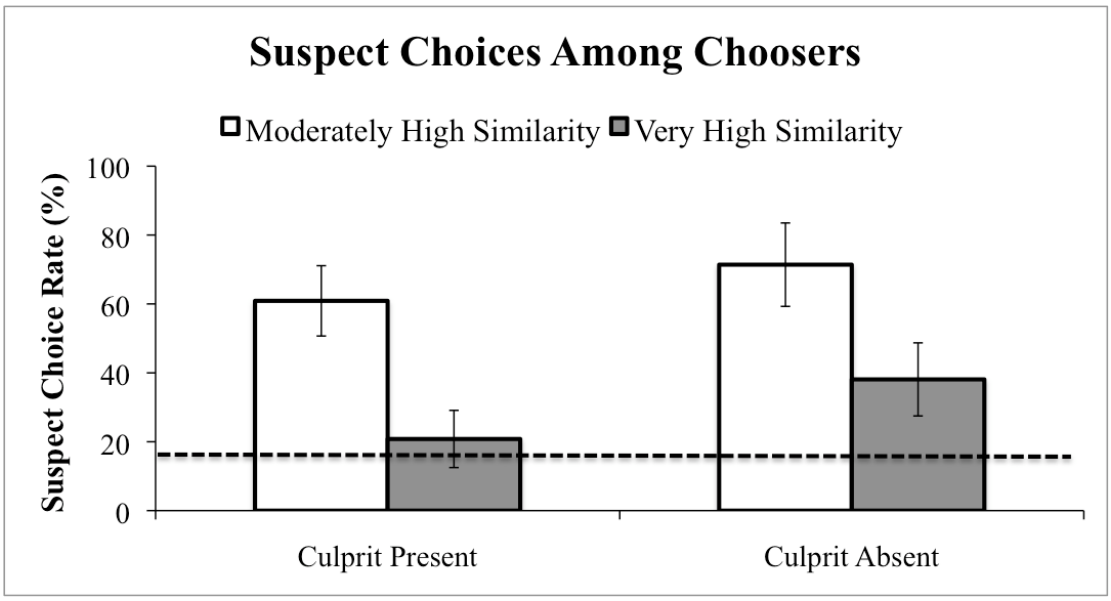

Figure 4. Proportion of suspect choices (culprit for culprit-present lineups; innocent suspect for culprit-absent lineups) from among only those who picked someone from the lineup (choosers). The dotted line indicates the suspect choice rate expected by chance. Error bars represent +/- 1 standard error.

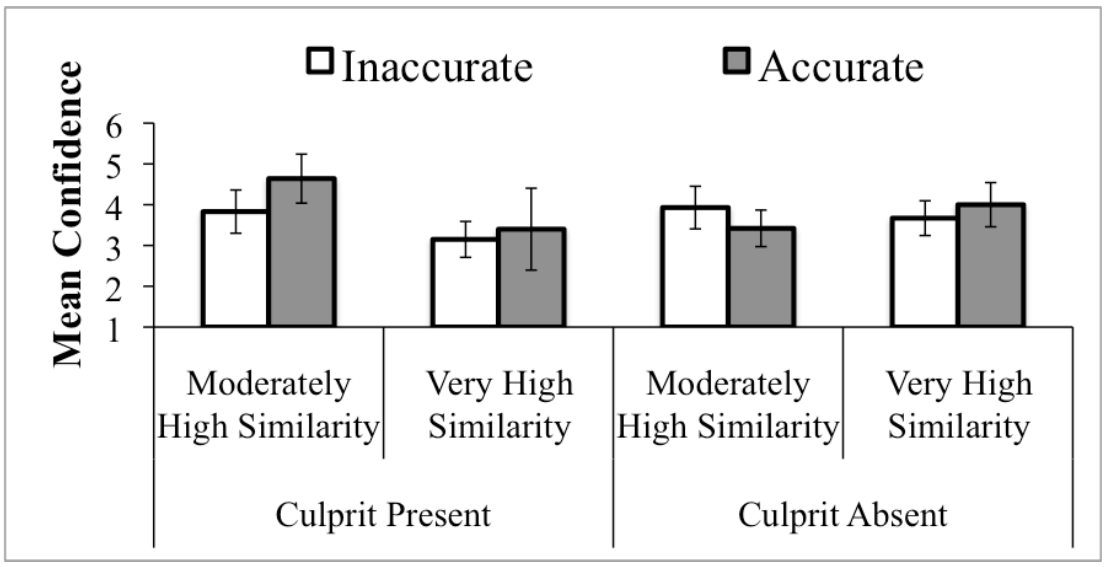

Figure 5. Post-identification confidence ratings as a function of culprit-presence and identification accuracy. The scale ranged from 1 (not at all confident) to 6 (very confident). Error bars represent 95% confidence intervals.