The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis

Ryan J. Fitzgerald[1], Heather L. Price[1], Chris Oriet[1], & Steve D. Charman[2]

[1]University of Regina

[2]Florida International University

This article has been accepted for publication in *Psychology, Public Policy, and Law*.

This research was supported by an Alexander Graham Bell Canada Graduate Scholarship (Doctoral) from the Natural Sciences and Engineering Research Council of Canada to the first author, grants from the Natural Sciences and Engineering Research Council of Canada to the second and third authors, and Canada Foundation for Innovation Leader's Opportunity Funds to the second and third authors.

Correspondence to: Ryan J. Fitzgerald, who is now at Department of Psychology, University of Portsmouth, King Henry Building, King Henry 1 Street, Portsmouth, United Kingdom, PO1 2DY. E-mail: ryan.fitzgerald@port.ac.uk

Abstract

Eyewitness lineups are typically composed of a suspect (guilty or innocent) and fillers (known innocents). Meta-analytic techniques were used to investigate the extent to which manipulations of suspect-filler similarity influenced identification decisions. Compared with lineups with moderate or high similarity fillers, lineups with low similarity fillers were far more likely to elicit suspect identifications. This was true regardless of whether the suspect was guilty or innocent, underscoring the importance of ensuring the suspect does not stand out from the fillers. Although whether the lineup contained moderate or high similarity fillers had no reliable influence on guilty suspect identifications, a higher rate of innocent suspect misidentifications was found for moderate similarity lineups. The correspondence between the meta-analytic findings and current lineup construction recommendations is discussed. Keywords: similarity, eyewitness identification, meta-analysis, filler, lineup composition

The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis

Numerous factors warrant consideration when constructing a lineup for eyewitness identification. One consideration is the degree to which fillers should resemble the police suspect. Appropriate selection of fillers is crucial for creating a fair lineup that balances the competing demands of minimizing false identification with maximizing culprit identification. A report published by an interdisciplinary panel of eyewitness experts recommends constructing lineups to ensure "the suspect does not unduly stand out" (Technical Working Group for Eyewitness Evidence, 2003, p. 32). Although this requirement could be met by using fillers who are highly similar in appearance to the suspect, in that same report investigators are further advised to "avoid using fillers that so closely resemble the suspect that a person familiar with the suspect might find it difficult to distinguish the suspect from the fillers" (p. 33). In essence, the suspect should be accompanied by fillers who are neither too dissimilar nor too similar.

The present meta-analysis was conducted to provide an overview of how suspect-filler similarity affects identification outcomes. A meta-analytic approach is helpful because the results of one study may, for example, be contingent upon the materials employed and might not generalize to all similarity manipulations (for a discussion of the need for stimulus sampling, see Wells & Windschitl, 1999). Although the number of studies with direct manipulations of suspect-filler similarity is relatively small, there have been several instances in which similarity has been indirectly manipulated (e.g., by adopting different filler selection strategies). This meta-analytic review summarizes studies that often differ in their stated intentions yet each contain a common element—a comparison between lineups that differ in suspect-filler similarity. Synthesizing data collected from different sources allows for a better understanding of how suspect-filler similarity affects lineup choices across different identification conditions. A

comprehensive understanding is especially desirable in the case of suspect-filler similarity manipulations, because researchers in this domain have yet to implement a standard method of objectively determining the similarity between two persons (Tredoux, 2002).

Filler Selection

A typical lineup is composed of a police suspect (who may be the culprit or may be an innocent suspect) and a set of fillers who are known to be innocent. The strategy employed to select fillers can influence the extent to which they resemble the suspect. Previous research has focused on two filler selection strategies: matching to the appearance of the suspect and matching to a description of the culprit.

Match to Appearance

The most commonly used method of selecting fillers is to match them to the appearance of the suspect (Wogalter, Malpass, & McQuiston, 2004). When this method is used, the critical concern is determining the appropriate level of similarity between the fillers and the suspect. Luus and Wells (1991) briefly mentioned the possibility of an optimal-similarity function in which the relation between suspect-filler similarity and lineup diagnosticity— that is, the likelihood that a suspect identification is of the culprit rather than an innocent suspect—would be characterized by an inverted U-shape. They suggested low similarity lineups would be expected to have low diagnosticity because they would induce false identifications, and high similarity lineups would also be expected to have low diagnosticity because having lineup members that look too similar to the suspect would impede correct identifications. Accordingly, an optimal lineup would include fillers who lie somewhere in the middle of the similarity spectrum.

Although fillers matched to the suspect were initially thought to protect innocent suspects from false identification (Lindsay & Wells, 1980), researchers have speculated that matching

fillers to the suspect's appearance could "backfire" and actually increase the likelihood of innocent suspect misidentification (Clark, 2003). The first to note this possibility was Navon (1992), who pointed out that suspects and fillers are placed in appearance-matched lineups for different reasons. Unlike the fillers, who are selected because of their match to the suspect, innocent suspects are often in a lineup because of their match to a description of the culprit. As a consequence, the innocent suspect would be the lineup member who is most similar to the perpetrator and thus would also be the most likely to be misidentified. Wogalter, Marwitz, and Leonard (1992) proposed an alternative mechanism by which appearance-matched lineups could lead to innocent suspect misidentifications. Their reasoning rests on the notion that because suspects are the origin of suspect-matched lineups, they will be more similar to the fillers than any of the fillers are to each other. Therefore, innocent suspects might be chosen because they represent the central tendency of the lineup. Consistent with these predictions, an innocent suspect was chosen from one appearance-matched lineup at a higher rate than all the fillers combined (Clark & Tunnicliff, 2001).

Match to Description

Luus and Wells (1991) recommended using fillers who fit the eyewitness description of the culprit, but who also possess additional features that differ from those of the culprit. They hypothesized that matching on features in the description would protect innocent suspects from false identification and that allowing features not mentioned in the description to vary would facilitate recognition. Wells, Rydell, and Seelau (1993) provided convincing evidence in support of this claim. Compared with lineups composed of fillers who were low in similarity to the culprit, matching fillers to the witness description resulted in a 30% reduction in false identifications of the innocent suspect and had virtually no effect on correct identifications of the

culprit. Moreover, compared with lineups composed of fillers high in similarity to the culprit, matching fillers to the description resulted in a 45% increase in correct identifications and an equivalent rate of false identifications. Thus, lineups with fillers of moderate similarity offered protection to innocent suspects without impeding culprit identifications.

The match-to-description advantage observed by Wells et al. (1993) provided empirical support for the theoretical framework proposed by Luus and Wells (1991). When eyewitnesses describe a culprit, they draw on recall memory. The function of a lineup is to give the witness an opportunity to provide new, recognition-based information. When the suspect matches the description and the fillers do not, witnesses need not rely on recognition memory because the fillers can be discounted based on the incongruence between their appearance and what was recalled of the culprit's appearance. If the suspect happens to be innocent, the witness will be prone to false identification because the suspect will stand out from the fillers. Conversely, when fillers match the description, witnesses are forced to rely on recognition because all lineup members correspond with their recall. This explains the match-to-description procedure's protection of innocent suspects from false identification.

The advantage of matching to the witness description over matching to the suspect's appearance is explained by the complementary concepts of propitious heterogeneity and gratuitous similarity (Wells, 1993). Lineups with fillers who are matched only to the features in the description promote propitious heterogeneity, which is the idea that variations in the features not mentioned in the eyewitness description aid the process of recognition. Lineups with fillers who are matched to the appearance of the suspect promote gratuitous similarity, which is the idea that matching fillers to features of the suspect that are above and beyond those provided in the

witness description eliminates important differences among the lineup members that are needed for recognition to operate effectively.

Although the match-to-description procedure showed early promise, it evinced little or no advantage over the match-to-appearance procedure in subsequent research. In two experiments, relative to appearance-matched lineups, description-matched lineups produced only nonsignificant increases in culprit identifications (Juslin, Olsson, & Winman, 1996; Lindsay, Martin, & Webber, 1994). Moreover, in one of those experiments (Lindsay et al., 1994), the innocent suspect misidentification rate was significantly higher in description-matched lineups than in appearance-matched lineups, an effect the authors attributed to witness descriptions that were too vague. In more recent experiments, filler selection strategy had no effect on culprit or innocent suspect choice rates (Darling, Valentine, & Memon, 2008; Tunnicliff & Clark, 2000). To explain the absence of a match-to-description benefit in culprit identifications, Tunnicliff and Clark suggested the fillers in their appearance-matched lineups might not have resembled the suspect to the same degree as those in the experiment by Wells et al. (1993). These findings suggest matching to the witness description is not always the most advantageous filler selection procedure.

Furthermore, the match-to-description procedure is not always a viable option. Luus and Wells (1991) outlined three situations in which the fillers should not be matched to the eyewitness description: (a) When the description does not correspond with the appearance of the suspect; (b) When the description is so specific that it would be impossible to find fillers who match it; and (c) When multiple eyewitnesses to the same event report descriptions that contradict one another. Another problem with the match-to-description procedure is that eyewitnesses often provide generic face descriptions, which can result in a lineup with fillers

who fit the description yet look nothing like the suspect (Koehnken, Malpass, & Wogalter, 1996). Even more problematic is the situation in which eyewitnesses provide no information about the culprit's face. In one study (Lindsay et al., 1994), fewer than 10% of witnesses mentioned facial features in their description of a confederate to whom they had spoken for 3 min. Another obvious problem with the match-to-description procedure is that the eyewitness description could be inaccurate (Meissner, Sporer, & Schooler, 2007; Wells et al., 1998). These issues could partly explain why only 9% of respondents in a survey of police investigators indicated using witness descriptions as a basis for selecting fillers (Wogalter et al., 2004).

Simultaneous Versus Sequential Presentation

The manner in which a lineup is presented could moderate the effect of suspect-filler similarity on identification outcomes. Lineup members can be presented to eyewitnesses all at once (simultaneously) or one at a time (sequentially). Simultaneous lineups have been criticized for allowing witnesses to adopt a relative judgment strategy (Wells, 1984). That is, witnesses viewing simultaneous lineups might be tempted to choose the person who looks most similar to the perpetrator in comparison to the other lineup members. Although using relative judgments should tend to lead to correct identifications from culprit-present lineups, this strategy is clearly problematic when the culprit is absent. Lindsay and Wells (1985) consequently developed the sequential lineup procedure to encourage witnesses to compare each individual lineup member with their memory of the criminal (i.e., an absolute judgment) rather than with the other lineup members.

In the sequential lineup, each photo is shown individually and the witness is required to decide whether the photo is or is not the culprit before proceeding to the next one. Lindsay and Wells (1985) recommend showing each photo only once and not allowing witnesses to go back

and look at previously viewed photos. In addition, to prevent the tendency to choose someone near the end of the array, they suggest withholding the number of photos that will be viewed from the witness (commonly referred to as "backloading"). When Lindsay and Wells directly compared the two lineup presentation formats, they observed a lower false identification rate on the sequential lineup (17%) than on the simultaneous lineup (43%). Recent meta-analyses (Steblay, Dysart, Fulero, & Lindsay, 2001; Steblay, Dysart, & Wells, 2011) have supported the view that sequential lineups offer an effective safeguard against false identification. However, correct identification rates also tend to be lower in sequential compared with simultaneous lineups, leading some researchers to suggest the sequential procedure encourages witness to adopt a more conservative decision criterion (Flowe & Bessemer, 2011; Flowe & Ebbesen, 2007; Meissner, Tredoux, Parker, & MacLin, 2005). In other words, witnesses might be less willing to choose from sequential lineups than from simultaneous lineups.

Carlson, Gronlund, and Clark (2008) hypothesized that suspect-filler similarity could moderate lineup presentation effects. Specifically, they predicted that a sequential advantage would only emerge when an innocent suspect resembles the culprit and stands out from other lineup members who do not (i.e., when the lineup is biased), arguing the innocent suspect is much less likely to stand out in a sequential lineup because comparisons among lineup members are more difficult. Carlson et al. conducted two experiments to test their hypothesis. In the first experiment, simultaneous and sequential lineups that only contained fillers highly similar to the innocent suspect were compared. In support of their hypothesis that the sequential advantage would only emerge when lineups were biased, the lineups with highly similar fillers led to equivalent rates of innocent suspect misidentifications between the simultaneous and sequential conditions. Furthermore, a simultaneous advantage was found for correct identifications from

culprit-present lineups; however, Carlson et al. noted that this finding was likely a consequence of matching the fillers to the innocent suspect's appearance—even in culprit-present lineups—which would be expected to make the culprit stand out when the lineup members are presented simultaneously.

In their second experiment, Carlson et al. (2008) directly manipulated lineup fairness by constructing lineups containing fillers of low, moderate, and high similarity to the suspect. Again, compared with sequential lineups, simultaneous lineups produced a higher rate of correct identification from culprit-present lineups, albeit only for those containing low similarity fillers (i.e., biased lineups). In addition, in culprit-absent lineups the false identification rate was lower when presented sequentially than when presented simultaneously; however, similar to the simultaneous advantage for culprit-present lineups, the sequential advantage was only found in the case of biased lineups. These results indicate the degree of similarity between fillers and the suspect can influence the effects of lineup presentation manipulations. As a consequence, we included lineup presentation as a moderator variable in the present research.

Meta-Analytic Approach

When examining suspect-filler similarity effects, one could contrast the rate at which culprits and innocent suspects are chosen between lineups with fillers of high versus low resemblance to the suspect. Essentially, this was the approach used by Clark and Godfrey (2009) in their broad review of the eyewitness literature. Based on a summary of seven studies, Clark and Godfrey concluded that both culprit and innocent suspect choices were more prevalent when suspect-filler similarity was low than when it was high. The increase in innocent suspect misidentifications was greater than the increase in culprit identifications, suggesting that a low similarity lineup is more likely to lead to innocent suspect misidentifications than it is to

facilitate culprit identifications. Clark and Godfrey noted, however, that three of the seven studies summarized contained a manipulation of clothing bias rather than a manipulation of facial appearance. Clark and Godfrey also did not comment on how suspect-filler similarity affected filler selections and lineup rejections, focusing instead on suspect identifications. Filler selections are known errors, so they are less concerning than innocent suspect misidentifications (Wells & Lindsay, 1980). However, witnesses may be required to attempt lineup identifications on multiple occasions (Behrman & Davey, 2001); if a filler had been misidentified on a previous occasion, the credibility of that witness could be jeopardized (Tunnicliff & Clark, 2000). Furthermore, filler identifications have been shown to have diagnostic value; in fact, when the a priori likelihood of the suspect's guilt is relatively high, filler identifications may be more informative of a suspect's innocence than suspect identifications are of a suspect's guilt (Wells & Olson, 2002).

In the present research, we used meta-analytic techniques to investigate the effect of suspect-filler similarity on all identification responses (suspect identifications, filler identifications, and lineup rejections). In addition, rather than strictly comparing high and low similarity lineups, both of which were recommended against by the Technical Working Group for Eyewitness Evidence (2003), we compared three levels of suspect-filler similarity: low, moderate, and high. Note that these labels correspond to relative differences in similarity, rather than to objectively defined categories. Compared with moderate similarity lineups, low similarity lineups were expected to increase both correct identifications and false identifications. Conversely, high similarity lineups were expected to decrease correct identifications and false identifications.

Method

Procedure

Literature search. To begin, one of the authors searched the *PsycInfo* and *Web of Science* databases for articles containing various combinations of the following search terms: *eyewitness, identification, lineup, foil, filler, distractor, similarity, match, appearance, description, construction, fairness, composition*. When a relevant article was identified, its reference list, as well as the list of articles in which it had been cited, were examined. Finally, *Google* and *Google Scholar* were searched to account for any articles that were not in the aforementioned databases. The search ended in April 2012 with 17 independent studies that met the inclusion criteria, providing data from 6,650 participants. Publication dates for the articles ranged between 1980 and 2011.

Inclusion criteria. To be included in the meta-analysis, the study needed to be an investigation of event memory (live events or video events) that included a comparison between two or more lineups that differed in suspect-filler similarity. Several methods of checking similarity manipulations were present in the literature. One method is to obtain similarity ratings between the suspect and each of the fillers. That is, raters observed two faces (the suspect and a filler) and indicated their judgment of similarity using a Likert scale. The number of points on the scale varied widely from study to study, with some as low as 4 points and others as high as 1,001 points. Similarity judgments were typically, but not always (e.g., Juslin et al., 1996), conducted using participants who did not provide data for the study itself. Another method of checking similarity involves conducting mock witness tests in which a set of judges view a lineup and identify the person who best fits an eyewitness description of the suspect (Doob & Kirshenbaum, 1973). The data obtained from mock witness tests can then be used to calculate

effective size scores (Malpass, 1981; Tredoux, 1998), which are estimates of the number of lineup members who fit the description sufficiently to draw choices away from the suspect. Lineups with high effective size scores are judged to have higher suspect-filler similarity than lineups with low effective size scores (Brigham & Brandt, 1992; Brigham, Ready, & Spier, 1990).

Exclusion criteria. Studies that manipulated lineup similarity within the context of face recognition paradigms (e.g., Flowe & Ebbesen, 2007) were excluded because they were not considered to adequately correspond with the experience of an eyewitness. Clothing bias manipulations were also excluded because they were considered fundamentally different from facial similarity manipulations. Furthermore, within-subject designs were excluded because we could not be certain that a repeated-measures design would be assessing the same effect as an independent-groups design (Morris & DeShon, 2002). In some studies, identifications were made from two lineups. For example, in one study a target-absent lineup was followed by a target-present lineup (Read, Tollestrup, Hammersley, McFadzen, & Christensen, 1990). These studies were dealt with by only including data from the first lineup that was shown. Following the procedure employed by previous meta-analysts in the psychology and law domain (Deffenbacher, Bornstein, Penrod, & McGorty, 2004), unpublished studies were excluded to accommodate the legal system's preference for published research (e.g., *Daubert v. Merrell Dow Pharmaceuticals*, 1993).

Assignment of study weights. Because of the relative nature of similarity judgments, variability in true effect sizes was assumed and study weights were assigned using the random-effects model (Hedges, 1992). In contrast to the fixed-effect model, which only takes the within-study variance into account (i.e., sampling error) and assigns substantially greater weight to

larger studies than to smaller studies, in the random-effects model the weights assigned to smaller and larger studies are more evenly distributed because both the within-study variance and the between-study variance are taken into account (Borenstein, Hedges, Higgins, & Rothstein, 2010). The weight (*W*) for a given study in the random-effects model is calculated as the inverse of the within-study variance (*V*) and the estimated between-study variance ($T2$) combined:

$$W = \frac{1}{V + T^2}$$

Coding. To account for the varying degrees of suspect-filler similarity, lineups were categorized as having low, moderate, or high similarity. Three of the study authors were involved in the coding process. One author developed a coding guide that provided a general description of the fillers, a range of mean similarity ratings, and a range of effective size scores for each lineup similarity category (descriptions of the categories as well as details about the coding guide are provided in Appendix A; The individual studies within each category, as well as the proportions of each identification response, are provided in Appendices B–D). The other two authors independently coded the lineups using information provided in the Methods section of each article. Cohen's kappa indicated the initial level of interrater reliability was acceptable, $\kappa = .87$. All coding discrepancies were resolved to consensus through discussion between the two coders.

Effect size. Typically, one of three choices is possible in lineup identification tasks: suspect identifications, filler identifications, or lineup rejections.[1] For the purpose of the present

[1] Researchers occasionally provided a "not sure" option. For ease of comparison among studies that did or did not provide this option, all "not sure" outcomes were treated as lineup rejections.

research, the outcomes of each of these possibilities were treated as binary (e.g., the suspect was identified or the suspect was not identified) and analyzed in separate meta-analyses. This approach involved a relatively high number of tests, and the increased likelihood of Type I errors should be noted. However, by analyzing each outcome, we were able to determine whether changes in the rates of suspect identifications corresponded with changes in filler identifications or changes in lineup rejections.

When dealing with binary data, meta-analysts have the option of computing an odds ratio, a risk ratio, or a risk difference. Of the three measures, odds ratio has the best mathematical properties. For example, risk ratio and risk difference are not typically able to assume their full range of values, but odds ratio is capable of assuming its full range of values (Fleiss & Berlin, 2009). However, odds ratio is also the least intuitive of the three measures (Deeks, 2002). Risk difference is based on raw units and can be easily interpreted by both researchers and professionals who are unfamiliar with statistical techniques. For example, if rates of false identification were 25% for low similarity lineups and 35% for moderate similarity lineups, then the risk difference would be 10%. A meta-analysis of previous meta-analyses indicated similar conclusions were reached regardless of whether odds ratio or risk difference was calculated (Engels, Schmid, Terrin, Olkin, & Lau, 2000). Nevertheless, we calculated both measures to ensure our results would be both accurate and accessible. *Z* tests were computed to test each effect size measure for statistical significance.

Diagnosticity. Diagnosticity ratios provide a measure of the probative value of a lineup procedure by indicating how much more likely a suspect identification is to correspond to the perpetrator, as opposed to an innocent person. Diagnosticity is calculated as the ratio of culprit identifications from culprit-present lineups to innocent suspect identifications from culprit-absent

lineups (Wells & Lindsay, 1980). For example, if a guilty suspect (culprit) is chosen by 60% of witnesses from a culprit-present lineup and an innocent suspect is chosen by 20% of witnesses from a culprit-absent lineup, the diagnosticity ratio would be 3.0 (60 _ 20), indicating that a guilty suspect is 3 times more likely to be chosen than an innocent suspect. We compared these ratios among lineups that differed in suspect-filler similarity to evaluate the extent to which lineup fillers affected diagnosticity. For diagnosticity to be compared, both similarity and culprit presence needed to be manipulated within the study. Because not all studies that met the inclusion criteria had both culprit-present and culprit-absent conditions, the diagnosticity statistics were based on a subset of the studies included in the meta-analysis. Furthermore, not all studies included low, moderate, and high similarity lineups. As a consequence, if a study included only a comparison between low and moderate similarity lineups, data from the moderate similarity lineup from that study would not affect the diagnosticity ratios that were calculated for the comparison between moderate and high similarity lineups. This approach was important because it allowed for causal conclusions regarding any differences in diagnosticity between two lineup types; however, the overall diagnosticity values for all the low, moderate, and high similarity lineups are provided in Appendix E.

Moderator Variables

Presentation. Given the different strategies that are hypothesized to play a role in evaluating simultaneous (relative judgment) and sequential (absolute judgment) lineups, lineup presentation could have an influence on suspect-filler similarity effects. Carlson et al. (2008) showed that suspect-filler similarity can moderate the effect of simultaneous versus sequential lineup presentation. In the present research, we took a different approach and tested whether

lineup presentation moderates the effect of manipulating suspect-filler similarity by including lineup presentation as a categorical moderator variable.

Culprit presence. Suspect-filler similarity has been manipulated in both culprit-present and culprit-absent lineups. In the case of culprit-present lineups, suspect-filler similarity always refers to the same thing: the degree to which the fillers resemble the culprit. For culprit-absent lineups, suspect-filler similarity could again refer to the degree of similarity between the culprit and the fillers or it could refer to the degree of similarity between the innocent suspect and the fillers. This discrepancy is a consequence of some researchers striving for experimental control by using the same fillers in culprit-present and culprit-absent lineups and other researchers striving for ecological validity by following the procedures that would be used when lineups are constructed by law enforcement personnel (Clark & Tunnicliff, 2001). For the present purposes, we made no distinction between culprit-absent lineups with fillers who were matched to the perpetrator and culprit-absent lineups with fillers who were matched to the innocent suspect, instead focusing on the similarity between lineups irrespective of the method used to manipulate it. Although it would be ideal to analyze the two variations of suspect-filler similarity on culprit-absent lineups separately, the limited number of studies examining suspect-filler similarity made this approach undesirable.

Whether the culprit is present or absent in a lineup necessarily determines the outcome that is considered accurate. For culprit-present lineups, the correct decision is to identify the suspect. For culprit-absent lineups, the correct decision is to reject the lineup. Given the fundamental difference between a lineup that contains a culprit and one that does not (Wells & Penrod, 2011), separate meta-analyses were conducted for culprit-present and culprit-absent lineups instead of including culprit-presence as a moderator within a larger meta-analysis.

Results

Two effect size measures were calculated for all analyses: odds ratio and risk difference. Consistent with previous research (Engels et al., 2000), none of the main effects differed in statistical significance as a function of the outcome measure. The only moderator analysis that differed by outcome measure was for filler identifications in the comparison between high and moderate similarity lineups when the culprit was absent. Specifically, a risk difference that was marginal (p = .056) corresponded with an odds ratio that was significant (p = .035). For the sake of avoiding redundancy, only risk difference (the more intuitive measure) is reported. The odds ratio analyses can be obtained by contacting the first author.

Risk Differences (Effect Size) and Diagnosticity

Table 1 presents descriptive and inferential statistics for the main effects of similarity. All analyses have been divided by whether the culprit was present or absent from the lineup. For each of the three comparisons (high vs. low similarity lineups, moderate vs. low similarity lineups, and moderate vs. high similarity lineups), the proportions of suspect identifications, filler identifications, and lineup rejections are provided, as well as the difference between those proportions (risk difference). The risk difference was computed such that a positive value would indicate that as similarity increased, so did the likelihood of a given outcome. Conversely, a negative risk difference would indicate that as similarity increased, the likelihood of a given outcome decreased. Table 1 also includes 95% confidence intervals, null hypothesis significance tests, and heterogeneity tests associated with the risk differences. Finally, Table 1 also includes the number of studies (*K*) and the number of participants (*N*) associated with each of the three comparisons.

High versus low similarity (Table 1). For culprit-present lineups, the correct identification rate was significantly lower for high similarity lineups compared with low similarity lineups. Correspondingly, the filler identification rate was significantly higher for high similarity lineups compared with low similarity lineups. The difference in incorrect rejections between high and low similarity lineups was not reliable. Thus, it appears that including fillers who are highly similar to the suspect has the effect of drawing choices away from that suspect and toward the highly similar fillers, rather than toward rejection of the lineup.

When the culprit was absent, the rate of innocent suspect misidentifications was significantly lower for high similarity lineups compared with low similarity lineups. The rate of filler identifications was significantly higher for high similarity lineups compared with low similarity lineups. There was no reliable difference in correct rejections between high and low similarity lineups. This pattern of results closely mirrors that found when the culprit was present. Thus, regardless of whether the culprit was present or absent, replacing low similarity fillers with high similarity fillers resulted in a decrease in suspect identifications and an increase in filler identifications.

The ratio of culprit identifications from culprit-present lineups to innocent suspect misidentifications from culprit-absent lineups was calculated for both high and low similarity lineups to assess their diagnosticity. This analysis showed that the diagnostic value of suspect choices from high similarity lineups (5.07) was 3.11 times greater than the diagnostic value of suspect choices from low similarity lineups (1.64). Thus, compared with low similarity lineups, suspect identifications from high similarity lineups were more likely to be accurate choices.

Moderate versus low similarity (Table 1). Comparisons between moderate and low similarity lineups produced results similar to those found when high and low similarity lineups

were compared. Specifically, compared with low similarity lineups, moderate similarity lineups produced a lower rate of suspect identifications, a higher rate of filler identifications, and had no reliable effect on lineup rejections. Again, this pattern was obtained both when the culprit was present and when the culprit was absent. The ratio of culprit identifications to innocent suspect misidentifications between the two lineup types revealed a diagnostic value of moderate similarity lineups (3.27) that was 1.47 times greater than the diagnostic value of low similarity lineups (2.22).

High versus moderate similarity (Table 1). In contrast to the previous two comparisons, the effect of manipulating whether fillers were highly similar or moderately similar to the suspect was dependent upon whether the culprit was present or absent. For culprit-present lineups, there were no significant differences between high and moderate similarity lineups (for culprit identifications, filler identifications, and lineup rejections). In contrast, for culprit-absent lineups, the innocent suspect misidentification rate was significantly lower when fillers were highly similar compared with when they were moderately similar. A concomitant increase in the filler identification rate on high similarity lineups compared with moderate similarity lineups was also observed. Consistent with the high-low and the moderate-low comparisons, the difference in correct rejections between high and moderate similarity lineups was not reliable. The ratio of culprit identifications to innocent suspect misidentifications indicated that as similarity increased, so did diagnosticity. Specifically, the diagnostic value of high similarity lineups (10.67) was 2.50 times greater than the diagnostic value of moderate similarity lineups (4.28).

Heterogeneity

The heterogeneity of effect sizes among studies was assessed by two metrics: Cochran's *Q* and *I*2. Cochran's *Q* tests the null hypothesis that the true effect size does not vary from study

to study (Cochran, 1954). A significant *Q* value that is greater than expected by chance (that is, greater than the degrees of freedom) indicates the absence of a common effect size. Although the *Q* statistic provides an indication of whether or not there is greater heterogeneity than expected by chance, it does not indicate the amount of heterogeneity that is present. In contrast, *I2* gives an indication of the extent of heterogeneity. Specifically, *I2* provides an estimate of the proportion of the observed differences in effect size that were because of variations in true effects, as opposed to sampling error (Higgins, Thompson, Deeks, & Altman, 2003).

Table 1 provides ample evidence of heterogeneity in effect sizes, indicating the assignment of weights using the random-effects model was appropriate. Many of the *Q* tests were significant, indicating the obtained summary effects were often based on individual effect sizes that were subject to dispersion. This was particularly true for suspect and filler identifications, which were almost always significant. In contrast, only one *Q* test for lineup rejections reached significance. The mean of all *I2* values that were computed was 49.8. Using the benchmarks proposed by Higgins et al. (2003), a mean near 50 would suggest a moderate amount of the dispersion in effect sizes was due to real differences in true effects; however, a few *I2* values of zero were observed in the lineup rejection analyses, suggesting any dispersion observed in these effects was likely a consequence of sampling error.

Moderator Analysis: Lineup Presentation

Table 2 presents descriptive and inferential statistics for the moderator effect of lineup presentation, including the rate of each lineup outcome as a function of whether the lineup was presented simultaneously or sequentially, their associated risk differences, and a significance test (*Q*) of the moderating effect.

For culprit-present lineups, three significant moderator effects of lineup presentation were observed. First, in the comparison between moderate and low similarity lineups, culprit identifications were moderated by whether the lineup was presented simultaneously or sequentially. For both simultaneous and sequential lineups, the culprit was more likely to be identified from a low similarity lineup than from a moderate similarity lineup; however, the increase in culprit identifications was larger for sequential lineups than for simultaneous lineups. Second, when presented simultaneously the culprit identification rate was higher in moderate similarity lineups than in high similarity lineups. In contrast, when presented sequentially culprit identifications were less likely to occur in moderate similarity lineups than in high similarity lineups. Third, for simultaneous lineups the filler identification rate for moderate similarity lineups was slightly lower than the filler identification rate for high similarity lineups. Conversely, for sequential lineups the filler identification rate for moderate similarity lineups was higher than the filler identification rate for high similarity lineups.

For culprit-absent lineups, none of the moderating effects of lineup presentation reached significance. However, given the limited data available, it is worth considering some notable trends. Table 2 shows that decreases in suspect-filler similarity were associated with greater increases in innocent suspect misidentifications from simultaneous lineups than from sequential lineups. For example, compared with moderate similarity lineups, low similarity lineups increased innocent suspect misidentifications by 19% when presented simultaneously compared with 12% when presented sequentially. Similarly, in the comparison with high similarity lineups, low similarity lineups increased innocent suspect misidentifications by 23% when presented simultaneously compared with 11% when presented sequentially. Regardless of how the lineup was presented, innocent suspect misidentifications were consistently more likely with low

similarity lineups than with moderate or high similarity lineups; however, it appears this effect may be partly mitigated by the use of sequential lineups.

Discussion

The meta-analysis revealed several key findings:

(a) Suspect identifications were more common from low similarity lineups than from moderate or high similarity lineups. This was true both for culprit identifications and for innocent suspect misidentifications.

(b) Filler identifications were more common from moderate and high similarity lineups than from low similarity lineups. This finding was unaffected by whether the culprit was present or absent.

(c) Suspect-filler similarity had no reliable effects on lineup rejections, regardless of whether the culprit was present or absent.

(d) Whether the lineup contained moderate or high similarity fillers had no reliable effect on culprit identifications; however, innocent suspects were significantly more likely to be misidentified from moderate similarity lineups than from high similarity lineups.

(e) Increases in suspect-filler similarity corresponded with increases in the degree to which suspect identifications were diagnostic of the suspect's guilt.

As expected, the presence of low similarity fillers was associated with an increased likelihood of suspect identifications. This pattern was essentially uninfluenced by whether or not the culprit was present. Moreover, whether low similarity lineups were compared with moderate or with high similarity lineups also had little consequence. Such a robust finding emphasizes the value of ensuring the suspect does not stand out in a lineup. When filler similarity increased, there was a shift from suspect identifications to filler identifications rather than to lineup

rejections. In other words, similarity manipulations had no reliable effect on whether a lineup member was chosen or not. Rather, the similarity of fillers only seemed to influence *which* lineup member was chosen. If fillers were dissimilar, the suspect was more likely to be chosen. If fillers were similar, a filler was more likely to be chosen.

Given that increasing similarity resulted in a shift from suspect to filler identifications regardless of whether the culprit was present or absent, the meta-analytic findings are consistent with Clark's (2012) assertion that policies designed to prevent innocent suspect misidentifications come at the cost of reducing correct identifications of the culprit. Clearly, culprits are more easily identified when fillers are dissimilar-looking than when they are similar-looking. However, the diagnosticity ratios indicated that any reduction in culprit identifications associated with increased filler similarity was outweighed by a more pronounced reduction in innocent suspect misidentifications. As similarity between the suspect and fillers increased, the diagnosticity of suspect identifications also consistently increased. Thus, although it would be misleading to suggest increasing suspect-filler similarity had no cost, the reduction in culprit identifications was lower in magnitude than the reduction in innocent suspect misidentifications.

Suspect and Filler Identifications

Wells (1984) theorized that simultaneous lineups encourage witnesses to adopt a relative judgment strategy in which lineup members are compared with one another and the person who best resembles the culprit is chosen. Wells (1993) provided compelling evidence in support of this claim by comparing identification responses on a culprit-present lineup to a lineup that was identical except the culprit had been removed without replacement. From the culprit-present lineup, the culprit was chosen by approximately half of the witnesses and fillers were chosen by one quarter of the witnesses. If witnesses were using an absolute judgment strategy, removing

the culprit would be expected to facilitate a shift from culprit identifications to lineup rejections and the filler identification rate should have been unchanged; however, that was not the case. On the contrary, removing the culprit resulted in a shift from culprit identifications to filler identifications, which more than doubled. More recently, using the removal-without-replacement procedure, Clark and Davey (2005) replicated the shift from culprit to filler identifications in simultaneous lineups. Interestingly, a similar shift was present when lineup members were presented sequentially, leading Clark and Davey to suggest relative decisions might also occur with sequential lineups.

The meta-analytic results provide further support for the notion that witnesses engage in a relative judgment strategy when making lineup decisions. In culprit-absent lineups, the innocent suspects who were chosen by researchers typically either fit the culprit's description (Clark & Tunnicliff, 2001; Juslin et al., 1996; Lindsay & Wells, 1980; Tredoux, Parker, & Nunez, 2007; Wells et al., 1993) or were highly similar to the culprit's appearance (Carlson et al., 2008; Darling et al., 2008; Gronlund, Carlson, Dailey, & Goodsell, 2009; Lindsay et al., 1991). Thus, when lineups were composed of fillers who did not resemble the culprit, the innocent suspect would have been the lineup member who best matched the culprit's appearance. If participants were using relative judgments, this strategy should have increased the number of innocent suspect choices from low similarity lineups. Conversely, increasing the similarity of the fillers to the suspect should have increased the number of filler identifications, as it would have increased the likelihood that one of the fillers would have best resembled the culprit. This is precisely the pattern of results that was revealed in the meta-analysis.

Lineup Rejections

In previous studies, manipulations of suspect-filler similarity have produced conflicting effects on lineup rejections. Increasing similarity between the suspect and fillers has been associated with increases in lineup rejections (e.g., Carlson et al., 2008; culprit-present, simultaneous lineups), decreases in lineup rejections (e.g., Lindsay et al., 1991, Experiment 3; culprit-present, simultaneous lineups), as well as having no effect on lineup rejections (Brewer & Wells, 2006; Charman, Wells, & Joy, 2011; Clark & Tunnicliff, 2001; Darling et al., 2008; Juslin et al., 1996; Lindsay et al., 1994; Lindsay & Wells, 1980; Tredoux et al., 2007; Wells et al., 1993). Tunnicliff and Clark (2000) explored one situation in which similarity seems likely to affect lineup rejections: when the lineup has been matched to the appearance of an innocent suspect who does not resemble the culprit. In two experiments, they found that lineup rejections were commonplace when none of the lineup members resembled the culprit. Tunnicliff and Clark further discussed what might happen when the innocent suspect and the culprit are similar in appearance. If a similar-looking innocent suspect were placed into a lineup with dissimilar fillers, the lineup would be biased because the innocent suspect would stand out. In this scenario, a false identification seems more likely than a lineup rejection. Nevertheless, that has not always been the case. For example, when Lindsay et al. (1991, Experiment 3) biased a culprit-absent lineup toward an innocent suspect, identification responses were split almost evenly between false identifications of the innocent suspect and correct rejections of the lineup (fillers were never chosen). Moreover, the correct rejection rate for the biased lineup (46%) was twice the correct rejection rate for another lineup containing fillers who did resemble the culprit (23%). These data, combined with those reported by Tunnicliff and Clark, indicate that lineups low in

suspect-filler similarity are more likely to be rejected than lineups of moderate or high suspect-filler similarity.

It seems reasonable to hypothesize that lineup rejections would be inversely related to suspect-filler similarity. A lineup composed of fillers who are highly similar to the culprit should draw more choices than a lineup composed of fillers who bear little resemblance to the culprit. As intuitive as this idea may be, data suggesting the opposite have been reported. For example, Carlson et al. (2008) found a higher correct rejection rate for moderate similarity lineups (47%) than for low similarity lineups (24%), although this pattern was only found for simultaneous lineups. For sequential lineups, a nonsignificant trend in the opposite direction was observed. Although the effect observed by Lindsay et al. (1991) was consistent for simultaneous and sequential lineups, the results reported by Carlson et al. suggest lineup presentation might influence whether suspect-filler similarity influences lineup rejections.

Our evaluation of the literature on the whole showed that similarity rarely had an effect on whether or not a lineup was rejected. The absence of an effect of similarity on rejections was perhaps the most consistent finding in the meta-analysis. Regardless of whether low and high, moderate and high, or low and moderate lineups were compared, similarity effects on rejection were both small in magnitude and nonsignificant. Why was the rate of lineup rejections unchanged by manipulations of suspect-filler similarity? One possibility is that increasing the similarity of fillers produces contradictory effects. In his WITNESS model, Clark (2003) hypothesized that two factors contribute to suspect and filler identifications: (a) the extent to which a given lineup member matches the witness's memory of the culprit (i.e., an absolute judgment) and (b) the difference in strength of the recognition experience between the lineup member who best matches the witness's memory of the culprit and the next-best alternative.

Therefore, increasing suspect-filler similarity could increase the likelihood that a lineup member will match the witness's memory of the culprit and thus exceed the criterion for a choice to be made while simultaneously decreasing the difference between the best match and the next-best match, in turn decreasing the witness's confidence that the best match is in fact the culprit. Were this to be the case, these competing effects could, as observed in the present research, result in no net change in rejection rates. Of course, the effects would only negate each other if they are similar in strength. An effect of similarity on rejections could be expected if one of these competing effects was stronger than the other, which would explain why similarity has sometimes been observed to influence rejections.

Limitations of the Meta-Analysis

There are, of course, limitations of the meta-analysis that should be noted. First, suspect-filler similarity was operationalized as the average similarity of the fillers to the suspect. Thus, a moderate similarity lineup could consist entirely of fillers who moderately resemble the suspect or it could consist of some combination of fillers of low, moderate, and high resemblance. Furthermore, similarity relations are not limited to the resemblance between the suspect and the fillers. Other similarity relations that could affect eyewitness accuracy include the similarity between the culprit's photo and the culprit's physical appearance, the similarity among the fillers, and the similarity between the culprit and the innocent suspect. For instance, the extent to which the innocent suspect resembles the culprit would almost certainly influence suspect-filler similarity effects and ideally would have been included as a moderator variable. Unfortunately, this was not an option because ratings of the similarity between the culprit and the innocent suspect were rarely reported (but see Clark & Tunnicliff, 2001).

Second, we excluded unpublished studies from the analysis to accommodate the preference for published research in the legal system. Significant effects are generally more likely to be published than nonsignificant effects, so including unpublished studies into the meta-analysis might have resulted in smaller effects.

Third, despite more than 30 years having passed since the first exploration of suspect-filler similarity effects, the literature in this domain is relatively small. As a consequence, the number of studies comprising appropriate tests of similarity limited the scope of the meta-analysis. For example, researchers varied in how they manipulated similarity and a larger database would have been needed to effectively examine whether the type of manipulation influenced similarity effects. On a related note, because not all studies included a manipulation of lineup presentation, our conclusions about the moderating effect of this variable are tentative. In future studies of suspect-filler similarity, we encourage researchers to include the full design (similarity × culprit-presence × lineup presentation) to increase our understanding of the relation between these three variables.

Lineup Construction Recommendations

In the report developed by the Technical Working Group for Eyewitness Evidence (2003), police investigators are advised that the suspected culprit should not stand out from the lineup members who are known to be innocent. The meta-analysis results provide support for this recommendation. Compared with lineups that had fillers of moderate or high suspect-filler similarity, the rate of innocent suspect misidentifications nearly doubled when lineups contained fillers of low suspect-filler similarity. The group's report further advises police investigators to ensure that fillers and the suspect are not too similar. The concern is that using extremely similar fillers will essentially result in a lineup of "clones" that would greatly diminish the likelihood

that a culprit will be correctly identified. However, our synthesis of the existing literature did not support this assertion. Although the not-too-similar rule has a solid theoretical foundation (Luus & Wells, 1991) that was soon after supported by empirical research (Wells et al., 1993), we found no reliable difference in correct identifications between lineups within the categories of high and moderate suspect-filler similarity.

Wells (1993) reported concern among some eyewitness researchers that choosing fillers with features that vary from those of the suspect could result in lineups with an unintended bias toward innocent suspects. The present research suggests their concern may have been justified. Innocent suspects were significantly more likely to be misidentified from lineups of moderate suspect-filler similarity compared with lineups of high suspect-filler similarity. This increase in innocent suspect misidentifications, taken together with the null effect in culprit identifications, suggests that either (a) the rule of ensuring lineup members are not too similar to the suspect does not improve performance on culprit-present lineups and may actually contribute to wrongful convictions or (b) the inability to obtain fillers who are truly of high resemblance to the suspect has led to an incongruity between theory and practice. In other words, although the rule to avoid highly similar fillers may be theoretically sound, finding such fillers in practice may be more difficult than had been anticipated.

Lest it appear that we are advocating the dismissal of a rule that has been deemed best practice in lineup identification procedures (Turtle, Lindsay, & Wells, 2003), it is critical to emphasize that the similarity categories were developed in relation to one another and that the "high" similarity lineups might not have had the degree of similarity that has been cautioned against. Inspection of the similarity ratings provided by the researchers suggests this might very well have been the case. Although the lineups we categorized as "high" had ratings higher than

those categorized as “moderate,” and researchers sought to create very high similarity lineups in many cases, the high similarity lineups rarely had mean similarity ratings that were above the midpoint of the scales that were used. The lineups were certainly not comprised of clones, but the relatively modest similarity ratings may also indicate a reluctance of those judging similarity to use the upper end of the scale (see Flowe & Ebbeson, 2007).

In any event, we recommend additional research to further refine our understanding of what constitutes a lineup of fillers who are “too similar.” If our findings are replicated in future studies with lineups in which suspect-filler similarity is unquestionably high, then it might be best to advise using the most similar fillers available. Such a recommendation would provide less ambiguity than the current recommendation of using fillers who are similar, but not too similar.

References

*Studies marked with an asterisk were included in the meta-analysis

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25,* 475–491.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97-111. DOI: 10.1002/jrsm.12

*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

*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

*Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior, 35,* 479-500. DOI: 10.1007/s10979-010-9261-1

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. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238-259. DOI: 10.1177/1745691612439584

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior, 29,* 151-172. DOI: 10.1007/s10979-005-2418-7

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., & Tunnicliff, J. L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior, 25,* 199-216.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101-129.

*Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). Improving the reliability of eyewitness identification: Putting context into context. *Journal of Applied Psychology, 72,* 629-637.

*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

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine, 21,* 1575-1600. DOI: 10.1002/sim.1188

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28,* 687-706. DOI: 10.1007/s10979-004-0565-x

Doob, A. N., & Kirshenbaum, H. (1973). Bias in police lineups: Partial remembering. *Journal of Police Science and Administration, 1,* 287-293.

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine, 19,* 1707-1728.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. M. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis,* 2nd ed. (pp. 237–253). New York: The Russell Sage Foundation.

Flowe, H., & Bessemer, A. (2011). The effect of target discriminability and criterion placement on accuracy rates in sequential and simultaneous target-present lineups. *Psychology, Crime, & Law, 17,* 587-610. DOI: 10.1080/10683160903397540

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

*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

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279–296. DOI: 10.2307/1165125

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. J. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557-560.

*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.

Koehnken, G., Malpass, R. S., & Wogalter, M. S. (1996). Forensic applications of lineup research. In S. L. Sporer, R. S. Malpass & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 205-231). Mahwah, NJ: Erlbaum.

*Lindsay, R. C. L., James, A. L., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology, 76,* 796-802.

*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.

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70,* 556-564.

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. DOI: 10.1007/BF01044945

Meissner, C. A., Sporer, S. L., & Schooler, J. W. (2007). Person descriptions as eyewitness evidence. In R. C. L. Lindsay, D. E. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people* (Vol. 2), Mahwah, NJ: Erlbaum.

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33,* 783–792. DOI: 10.3758/BF03193074

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105-125. DOI: 10.1037//1082-989X.7.1.105

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

*Read, J. D., Tollestrup, P., Hammersley, R., McFadzen, E., & Christensen, A. (1990). The unconscious transference effect: Are bystanders ever misidentified? *Applied Cognitive Psychology, 4,* 3-31.

Steblay, N. K., Dysart, J. E., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25,* 459–473. DOI:10.1023/A:1012888715007

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

Technical Working Group for Eyewitness Evidence. (2003). *Eyewitness evidence: A trainer's manual for law enforcement*. Washington, DC: National Institute of Justice.

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22,* 217–237. DOI: 10.1023/A:1025746220886

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

*Tredoux, C. G., Parker, J. F., & Nunez, D.T. (2007). Predicting eyewitness identification accuracy with mock witness measures of lineup fairness: Quality of encoding interacts with lineup format. *South African Journal of Psychology, 37,* 207-222.

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

Turtle, J., Linsday, R. C. L., & Wells, G. L. (2003). Best practice recommendations for eyewitness evidence procedures: New ideas for the oldest way to solve a case. *Canadian Journal of Police and Security Services, 1,* 5-18

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14,* 89-103. DOI: 10.1111

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48,* 553–571. DOI: 10.1037//0003-066X.48.5.553

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88,* 776-784.

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviours. *Journal of Experimental Psychology: Applied, 8,* 155-167. DOI:10.1037//1076-898X.8.3.155

Wells, G. L., & Penrod, S. D. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. In B. Rosenfeld, & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 237-256). John Wiley and Sons, Hoboken, NJ.

*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,* 1-38.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115-1125.

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: 10.1080/106831604100016418

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

Table 1

Main effects of suspect-filler similarity on identification choices

						Effect Size and CI^{95}			Test of Null		Heterogeneity Indices		
Culprit	Lineup Choice	*K*	*N*	Similarity		Risk Diff.	Lower Limit	Upper Limit	*z*	*p*	*Q* (*df*)	*p*	I^2
				High	Low								
Present	Suspect	5	1338	.44	.65	-.21	-.34	.08	-3.16	.002	17.8 (4)	.001	77.5
	Filler	3	1138	.27	.06	.21	.14	.28	6.00	.001	3.7 (2)	.158	45.8
	Rejection	3	1138	.37	.34	.03	-.03	.08	1.02	.307	1.8 (2)	.411	00.0
Absent	Suspect	7	1775	.19	.37	-.18	-.27	-.10	-4.15	.001	23.2 (6)	.001	74.2
	Filler	9	1965	.36	.18	.18	.11	.25	5.05	.001	26.1 (8)	.001	69.4
	Rejection	9	1965	.47	.49	-.02	-.09	.05	-0.67	.505	15.7 (8)	.046	49.2
				Mod.	Low								
Present	Suspect	6	1449	.47	.63	-.16	-.26	-.05	-2.96	.003	15.0 (5)	.011	66.6
	Filler	5	1389	.24	.08	.16	.03	.29	2.42	.016	34.8 (4)	.001	88.5
	Rejection	5	1389	.32	.34	-.02	-.07	.03	-0.68	.499	1.8 (4)	.770	00.0
Absent	Suspect	7	1551	.24	.40	-.16	-.23	-.09	-4.40	.001	11.6 (6)	.072	48.3
	Filler	8	1583	.22	.10	.12	.06	.18	4.08	.001	13.9 (7)	.053	49.6
	Rejection	8	1583	.50	.48	.02	-.03	.07	0.81	.417	3.8 (7)	.804	00.0
				High	Mod.								
Present	Suspect	8	1996	.45	.47	-.02	-.12	.08	-0.43	.661	29.2 (7)	.001	76.1
	Filler	7	1938	.19	.17	.02	-.06	.10	0.47	.638	29.7 (6)	.001	79.8
	Rejection	7	1938	.37	.38	-.01	-.06	.04	-0.30	.762	7.1 (6)	.314	15.1
Absent	Suspect	9	1798	.12	.20	-.08	-.14	-.03	-2.82	.005	19.7 (8)	.012	59.4
	Filler	11	2425	.29	.22	.07	.01	.13	2.24	.025	25.9 (10)	.004	61.4
	Rejection	11	2425	.57	.58	-.01	-.06	.05	-0.27	.788	15.7 (10)	.110	36.1

Table 2

Moderating effects of lineup presentation on suspect-filler similarity manipulations

Culprit	Lineup Choice	Presentation	*K*	Similarity		Risk Difference	Moderator Test *Q*	*df*	*p*
				High	Low				
Present	Suspect	Simultaneous	5	.45	.69	-.24	0.5	1	.495
		Sequential	2	.35	.51	-.16			
	Filler	Simultaneous	3	.31	.07	.24	1.4	1	.237
		Sequential	2	.24	.06	.18			
	Rejection	Simultaneous	3	.36	.26	.10	0.8	1	.369
		Sequential	2	.42	.44	-.02			
Absent	Suspect	Simultaneous	7	.21	.44	-.23	2.6	1	.104
		Sequential	5	.16	.27	-.11			
	Filler	Simultaneous	9	.43	.20	.23	2.1	1	.149
		Sequential	5	.20	.07	.13			
	Rejection	Simultaneous	9	.38	.42	-.04	0.1	1	.759
		Sequential	5	.64	.66	-.02			
				Mod.	Low				
Present	Suspect	Simultaneous	6	.51	.66	-.15	5.8	1	.016
		Sequential	2	.22	.52	-.30			
	Filler	Simultaneous	5	.24	.08	.16	2.2	1	.137
		Sequential	2	.34	.06	.28			
	Rejection	Simultaneous	5	.26	.29	-.03	0.6	1	.427
		Sequential	2	.42	.41	.01			
Absent	Suspect	Simultaneous	7	.26	.45	-.19	0.7	1	.393
		Sequential	4	.21	.33	-.12			
	Filler	Simultaneous	8	.24	.11	.13	0.5	1	.477
		Sequential	4	.17	.07	.10			
	Rejection	Simultaneous	8	.47	.43	.04	0.1	1	.748

		Sequential	4	.61	.59	.02		
				High	Mod.			
Present	Suspect	Simultaneous	7	.42	.49	-.07	5.0	1	.022
		Sequential	2	.33	.23	.10		
	Filler	Simultaneous	6	.23	.20	.03	5.1	1	.024
		Sequential	2	.24	.35	-.11		
	Rejection	Simultaneous	6	.36	.35	.01	0.1	1	.743
		Sequential	2	.42	.45	-.03		
Absent	Suspect	Simultaneous	8	.13	.21	-.08	1.3	1	.258
		Sequential	4	.16	.19	-.03		
	Filler	Simultaneous	10	.37	.25	.12	3.7	1	.056
		Sequential	4	.21	.19	.02		
	Rejection	Simultaneous	10	.52	.55	-.03	0.6	1	.454
		Sequential	4	.62	.61	.01		

Appendix A

Coding Guide

Before the lineups were coded, guidelines were developed to facilitate a reliable method of categorizing lineups as having low, moderate, or high suspect-filler similarity. One of the authors developed the guidelines by reading the Methods sections of relevant articles and taking note of how labels researchers assigned to lineups (e.g., low similarity, high similarity) corresponded with similarity ratings and effective size values. However, coders were instructed to pay attention to more than just the similarity ratings and the effective size scores because these values could be influenced by the scale that was used (e.g., 7-point scale vs. 100-point scale) as well as the instructions that the researchers provided. Moreover, quantitative measures of similarity were not reported in some articles, further necessitating a more comprehensive approach to categorizing the lineups. To encourage the coders to consider more than the quantitative ratings and adopt a more holistic evaluation of the lineups, the guidelines included a range of mean similarity ratings for each of the categories that overlapped with one another. This provided coders with the flexibility to use information in addition to the similarity ratings and effective size scores when assigning a code to a lineup.

The range of similarity ratings and effective size scores for each category is provided below. All similarity ratings were converted to a 100-point scale for ease of comparison between studies. The lower end of the range of similarity ratings for the high similarity lineups may seem low on an absolute scale; however, there is good reason to suspect judges tend to be conservative when assessing the similarity between two faces. For instance, when Flowe and Ebbeson (2007) collected judgments of similarity between two computer-generated faces that apart from one feature were identical, those faces were assigned a similarity rating of 70 (on a 101-point scale).

Furthermore, of all the lineups included in the meta-analysis, not a single one exceeded a similarity rating of 60 (on a 101-point scale). Thus, in relative terms, a lineup with similarity ratings near the midpoint of a scale can be considered quite high. With regard to the effective size scores, our guidelines correspond well with Brigham (1990), who suggested a lineup with an effective size of 3 (for a 6-member lineup) should be considered "fair".

Low Similarity

Fillers bear little resemblance to the suspect. Similarity ratings range between 0 and 35 (on a 101-point scale). Effective size scores around 1-2 (for a 6-member lineup).

Moderate Similarity

Fillers resemble the suspect to some degree, but not as much as other potential fillers. Similarity ratings range between 25 and 50 (on a 101-point scale). Effective size scores around 2-3 (for a 6-member lineup).

High Similarity

Fillers closely resemble the suspect. Similarity ratings range between 40 and 100 (on a 101-point scale). Effective size scores around 4-5 (for a 6-member lineup).

Appendix B

Proportions of identification choices in high and low similarity lineups

		Culprit-Present						Culprit-Absent					
		Suspect		Filler		Rejection		Suspect		Filler		Rejection	
Study	Lineup	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
Cutler et al. (1987)	Sim	.65	.63	NR	NR	NR	NR	-	-	.73	.70	.27	.30
Read et al. (1990)	Sim	-	-	-	-	-	-	.14	.18	.27	.27	.59	.55
Lindsay et al. (1991) Exp 3	Sim	-	-	-	-	-	-	.40	.53	.37	.00	.23	.47
	Seq	-	-	-	-	-	-	.07	.07	.26	.00	.67	.93
Wells et al. (1993)	Sim	.21	.71	.43	.07	.36	.21	.12	.43	.48	.12	.41	.45
Lindsay et al. (1994) Exp 2	Sim	.66	.81	NR	NR	NR	NR	-	-	-	-	-	-
Lindsay et al. (1994) Exp 3	Sim	-	-	-	-	-	-	.08	.50	.39	.00	.53	.50
	Seq	-	-	-	-	-	-	.00	.16	.11	.00	.88	.84
Tredoux et al. (2007)	Sim	-	-	-	-	-	-	.25	.42	.34	.10	.41	.49
	Seq	-	-	-	-	-	-	.21	.41	.14	.15	.65	.44
Carlson et al. (2008)	Sim	.31	.71	.22	.06	.47	.24	.16	.64	.51	.12	.33	.24
	Seq	.41	.46	.20	.02	.39	.52	.20	.33	.16	.09	.64	.59
Gronlund et al. (2009)	Sim	.42	.62	.31	.07	.27	.31	.36	.47	.32	.14	.32	.39
	Seq	.31	.54	.25	.07	.44	.39	.28	.40	.28	.08	.44	.52
Charman et al. (2011)	Sim	-	-	-	-	-	-	-	-	.50	.38	.50	.63

Note: NR = Not Reported; Sim = Simultaneous Presentation; Seq = Sequential Presentation

Appendix C

Proportions of identification choices in moderate and low similarity lineups

		Culprit-Present						Culprit-Absent					
		Suspect		Filler		Rejection		Suspect		Filler		Rejection	
Study	Lineup	Mod	Low	Mod	Low	Mod	Low	Mod	Low	Mod	Low	Mod	Low
Lindsay et al. (1980)	Sim	.58	.71	.29	.12	.13	.18	.31	.70	.41	.04	.28	.26
Wells et al. (1993)	Sim	.67	.71	.07	.07	.26	.21	.12	.43	.31	.12	.57	.45
Lindsay et al. (1994) Exp 2	Sim	.79	.81	NR	NR	NR	NR	-	-	-	-	-	-
Lindsay et al. (1994) Exp 3	Sim	-	-	-	-	-	-	.25	.50	.19	.00	.56	.50
	Seq	-	-	-	-	-	-	.03	.16	.10	.00	.87	.84
Juslin et al. (1996)	Sim	.44	.52	.20	.11	.35	.38	.09	.09	.17	.12	.73	.78
Tredoux et al. (2007)	Sim	-	-	-	-	-	-	.42	.42	.12	.10	.46	.49
	Seq	-	-	-	-	-	-	.24	.41	.13	.15	.62	.44
Carlson et al. (2008)	Sim	.43	.71	.26	.06	.32	.24	.30	.64	.23	.12	.47	.24
	Seq	.24	.46	.24	.02	.53	.52	.38	.33	.17	.09	.46	.59
Gronlund et al. (2009)	Sim	.37	.62	.40	.07	.23	.31	.37	.47	.27	.14	.36	.39
	Seq	.22	.54	.39	.07	.39	.39	.23	.40	.25	.08	.52	.52
Charman et al. (2011)	Sim	-	-	-	-	-	-	-	-	.40	.36	.60	.64

Note: NR = Not Reported; Mod = Moderate; Sim = Simultaneous Presentation; Seq = Sequential Presentation

Appendix D

Proportions of identification choices in high and moderate similarity lineups

		Culprit-Present						Culprit-Absent					
		Suspect		Filler		Rejection		Suspect		Filler		Rejection	
Study	Lineup	High	Mod	High	Mod	High	Mod	High	Mod	High	Mod	High	Mod
Lindsay et al. (1991) Exp 1	Sim	.77	.67	.03	.03	.20	.30	.03	.20	.03	.10	.93	.70
Wells et al. (1993)	Sim	.21	.67	.43	.07	.36	.26	.12	.12	.48	.31	.41	.57
Lindsay et al. (1994) Exp 2	Sim	.66	.79	NR	NR	NR	NR	-	-	-	-	-	-
Lindsay et al. (1994) Exp 3	Sim	-	-	-	-	-	-	.08	.25	.39	.19	.53	.56
	Seq	-	-	-	-	-	-	.00	.03	.11	.10	.88	.87
Tunnicliff & Clark (2000)	Sim	.53	.53	.25	.16	.22	.31	.03	.13	.31	.34	.66	.53
Clark & Tunnicliff (2001)	Sim	-	-	-	-	-	-	.05	.25	.50	.16	.45	.59
Brewer & Wells (2006)	Sim	.40	.34	.18	.17	.42	.49	-	-	.33	.33	.67	.67
Darling et al. (2007)	Sim	.49	.45	.06	.09	.45	.47	.04	.05	.16	.21	.81	.74
Tredoux et al. (2007)	Sim	-	-	-	-	-	-	.25	.42	.34	.12	.41	.46
	Seq	-	-	-	-	-	-	.21	.24	.14	.13	.65	.62
Carlson et al. (2008)	Sim	.31	.43	.22	.26	.47	.32	.16	.30	.51	.23	.33	.47
	Seq	.41	.24	.20	.24	.39	.53	.20	.38	.16	.17	.64	.46
Gronlund et al. (2009)	Sim	.42	.37	.31	.40	.27	.23	.36	.37	.32	.27	.32	.36
	Seq	.31	.22	.25	.39	.44	.39	.28	.23	.28	.25	.44	.52
Charman et al. (2011)	Sim	-	-	-	-	-	-	-	-	.50	.40	.50	.60

Note: NR = Not Reported; Mod = Moderate; Sim = Simultaneous Presentation; Seq = Sequential Presentation

Appendix E

Diagnosticity ratios for lineups of high, moderate, and low suspect-filler similarity

Similarity	Study	Culprit IDs	Innocent Suspect IDs	Diagnosticity
High	Lindsay et al. (1991)	.767	.033	23.2
	Wells et al. (1993)	.214	.119	1.8
	Lindsay et al. (1994)	.660	.043	15.4
	Tunnicliff & Clark (2000)	.531	.031	17.1
	Darling et al. (2007)	.491	.035	14.0
	Carlson et al. (2008)	.360	.182	2.0
	Gronlund et al. (2009)	.364	.321	1.1
				M = 10.7
Moderate	Lindsay & Wells (1980)	.580	.410	1.4
	Lindsay et al. (1991)	.667	.200	3.3
	Wells et al. (1993)	.666	.119	5.6
	Lindsay et al. (1994)	.790	.149	5.3
	Juslin et al. (1996)	.200	.170	1.2
	Tunnicliff & Clark (2000)	.531	.125	4.2
	Darling et al. (2007)	.447	.047	9.5
	Carlson et al. (2008)	.327	.333	1.0
	Gronlund et al. (2009)	.296	.301	1.0
				M = 3.6
Low	Lindsay & Wells (1980)	.710	.700	1.0
	Wells et al. (1993)	.714	.429	1.7
	Lindsay et al. (1994)	.810	.338	2.4

Juslin et al. (1996)	.520	.090	5.8
Carlson et al. (2008)	.583	.505	1.2
Gronlund et al. (2009)	.579	.436	1.3
			M = 2.2