"This is an Accepted Manuscript of an article published by Taylor & Francis in Psychology, Crime, and Law on February 15, 2021, available online: http://www.tandfonline.com/[Article DOI]."

Validity of Mock-witness Measures for assessing Lineup Fairness

Jungwon Lee

Hallym University

Jamal K. Mansour

Queen Margaret University

Steven D. Penrod

John Jay College of Criminal Justice, The City University of New York

Author Note

Jungwon Lee, Department of Psychology and Hallym Applied Psychology Institute, Hallym University; Jamal K. Mansour, Memory Research Group, Centre for Applied Social Sciences, Psychology, Sociology, Education, Queen Margaret University; Steven D. Penrod, Department of Psychology, John Jay College of Criminal Justice, The City University of New York.

The authors declare no conflict of interest.

The present study is based upon work supported by the National Science Foundation under Grant No. SES-1754079. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Jungwon Lee.

Contact: jwl@hallym.ac.kr

Abstract

Although eyewitness researchers have used mock-witness measures to assess aspects of lineup fairness, they have paid little attention to their validity. The current study tested predictive validity, convergent validity, and discriminant validity of mock-witness measures from a meta-analytic perspective. Overall, mock-witness measures had predictive validity, particularly in target-absent (TA) lineups—the lineup fairness estimated by the measures reliably predicted eyewitnesses' choosing behaviors and discriminability of a suspect from fillers in TA lineups. However, correlations between lineup fairness estimated by mock-witnesses and eyewitness performance were significant in target-present (TP) lineups only when eyewitnesses had a moderate memory for the perpetrator. Multitrait-multimethod correlations demonstrated significant intradomain correlations between mock-witness measures and other lineup fairness indices and nonsignificant interdomain correlations between the mock-witness measures and indices reflecting memory strength for the perpetrator, which supported convergent validity and discriminant validity, respectively. The implications for research and practice are discussed.

Keywords: mock-witness, mock-witness paradigm, eyewitness, lineup fairness, lineup bias

Validity of Mock-witness Measures for Assessing Lineup Fairness

Prior studies have found that eyewitness performance is sensitive to characteristics of lineup composition, such as suspect presence (Clark, Howell, & Davey, 2008; Clark & Tunnicliff, 2001; Wells & Turtle, 1986), suspect position (Carlson, Gronlund, & Clark, 2008; Clark & Davey, 2005; Gronlund, Carlson, Dailey, & Goodsell, 2009), nominal lineup size (Cutler, Penrod, & Martens, 1987; Nosworthy & Lindsay, 1990; Pozzulo, Dempsey, & Wells, 2010), and lineup fairness (Bruer, Fitzgerald, Therrien, & Price, 2015; Colloff, Wade, & Strange, 2016).

Lineup fairness is a multi-faceted construct which has been loosely operationalized in the law. In Stovall v. Denno, 388 U. S. 293 (1967), the Court stated that the defendant could claim that "the confrontation conducted…was so unnecessarily suggestive and conducive to irreparable mistaken identification that he was denied due process of law." Stemming in part from the suggestiveness, researchers have regarded a lineup as unfair when the lineup structure implies that a particular lineup member is the suspect. Therefore, the composition of fair lineups focuses on "avoiding distinctiveness in the suspect's appearance which would indicate his identity to a witness" (Malpass & Devine, 1983, p.81). Eyewitness researchers have focused on two dimensions to compose fair lineups—the degree of similarity between target and fillers, and the number of reasonable fillers (Malpass, 1981). Thus, in the eyewitness literature, lineup fairness is conventionally defined as the extent to which a target does not stand out to someone with no prior experience of the crime's culprit and where the number of potential choices provides a sufficient range of alternatives that a guess has a probability of landing on the suspect at a rate that the courts would consider acceptable.

Prior studies have underscored that when fillers do not bear a resemblance to the suspect in a lineup (and particularly when an innocent suspect closely resembles the guilty suspect) eyewitnesses are more likely to identify the suspect regardless of whether that suspect is guilty or innocent, which may contribute to the wrongful conviction of an innocent person (e.g., Colloff et al., 2016; Clark & Tunnicliff, 2001). Recent studies suggest eyewitness performance can decrease in lineups where suspect-filler similarity is extremely high or extremely low (Bergold & Heaton, 2018; Carlson et al., 2019; Fitzgerald, Oriet, & Price, 2015; Lee & Penrod, 2019). For example, Fitzgerald et al. (2015) demonstrated that correct identifications were lower in very high-similarity lineups compared to moderately similar lineups. Lee and Penrod (2019) proposed in their multi-d′ model that eyewitnesses' discriminability of a guilty suspect from an innocent suspect increases as fillers more closely resemble the suspect, but there is a tipping point beyond which increases in filler similarity reduce eyewitness discriminability. Therefore, researchers have emphasized the importance of constructing lineups to improve the probative value of eyewitness identification.

However, it is challenging to assess and manage the similarity between a suspect and fillers in lineups in advance of presenting them to a witness. This is especially problematic for police who know what their suspect looks like and may have a more or less detailed description of the perpetrator but cannot know exactly what the perpetrator looks like. Even experimental researchers, who know precisely what their perpetrator looks likes, confront the task of assembling lineups which achieve some level of the suspect-filler similarity before they begin data collection from their eyewitnesses.

To measure lineup fairness in a laboratory, researchers have used the mock-witness paradigm (Doob & Kirshenbaum, 1973). In this paradigm, a description of a perpetrator is first

obtained. Commonly, a group of participants generate a description by viewing the perpetrator's face and writing descriptions of his or her appearance. Researchers then combine their descriptions and select features that are commonly mentioned by all or a proportion (e.g., 25% or 50%) of the participants as a modal description. Some researchers use the individual descriptions without combining them while others combine all features of a description. Another group of participants (i.e., mock witnesses), who have not seen the perpetrator previously, are given one of these descriptions and asked to choose a person in a lineup. The question asked varies but is generally along the lines of indicating who is most likely to be the perpetrator. Unlike eyewitnesses' task, mock witnesses are required to choose a lineup member—they typically cannot indicate 'not present' and they are not given a warning about a possible absence of the perpetrator in a lineup. From the identification performance of the mock witnesses, various quantitative indices measuring lineup fairness are calculated, such as Effective Size (Malpass, 1981), Tredoux's *E* (Tredoux, 1998, 1999), Acceptable Lineup Members (ALM; Malpass & Lindsay, 1999), Functional Size (Wells, Leippe, & Ostrom, 1979), and Suspect Bias (Doob & Kirshenbaum, 1973)—in the current study, we refer to these indices as mock-witness measures. Since researchers began to use the mock-witness paradigm in the 1970's, the use of the paradigm in research has become increasingly common. We found 259 journal articles in Google Scholar using ["mock-witness" paradigm] as a key word. Of the 259 articles, only 1 was published in the 1970 and only 2 were published in the 1980's; 16 articles were published in the 1990's; 80 articles were published in the 2000's; and 160 articles were published from 2010 to current.

Validity of the Mock-Witness Measures

Despite the increasing use of the mock-witness paradigm, only a few studies have examined its validity. Validity is referred to as the extent to which relevant evidence supports an

inference as being true or correct; and can be classified into four subtypes—statistical conclusion, internal, construct, and external (Shadish, Cook, & Campbell, 2002). Most studies testing the validity of mock-witness measures focused on construct validity (i.e., whether estimates of lineup fairness measured by the mock-witness paradigm represent conceptual constructs of lineup fairness). Construct validity itself has six subtypes—face, content, concurrent, predictive, convergent, and discriminant.

Face validity

Face validity is whether a measure appears to reflect what it is intended to measure. To assess face validity, untrained individuals or experts who have knowledge of the domain review items and make subjective judgments of the validity (Litwin, 1995). Since the validity of mock-witness measures has normally been tested using laboratory experiments, rather than using subjective reviews or ratings, there are no prior studies on the face validity of mock-witness measures. However, some researchers have questioned the face validity of the mock-witness paradigm itself by pointing out a lack of parallelism between eyewitnesses and mock witnesses. Corey, Malpass, and McQuiston (1999) proposed that the performance of mock witnesses may not be comparable to that of eyewitnesses because of differences in the cognitive processes used. That is, eyewitnesses identify a suspect by matching their internal visual representation of the perpetrator with the presented visual information whereas mock witnesses identify a suspect by matching written descriptions of the perpetrator with the visual information. However, the approach could be said to have some face validity by virtue of the fact that the response required from mock witnesses and eyewitnesses is identical or nearly so (i.e., selection of a lineup member).

Content validity

Although face validity and content validity are often used interchangeably, they are conceptually different. Content validity concerns the degree to which a measure represents a comprehensive sample of the theoretical content domain of a construct (Nunnally & Bernstein, 1994). Malpass (1981) considered what the content domain of lineup fairness comprises. He argued that lineup fairness comprises two dimensions—lineup size and lineup bias. Lineup size indicates the number of plausible members in the lineup, whereas lineup bias indicates the extent of which a suspect stands out from fillers in the lineup. According to Mansour, Beaudry, Kalmet, Bertrand, and Lindsay's (2017) classification, measures of lineup size include Effective Size, Tredoux's E, and ALM, while lineup bias includes Proportion of Suspect Selections (Brigham & Brandt, 1992; Doob & Kirshenbaum, 1973), Functional Size, Suspect Bias, Defendant Bias (Malpass, 1981; Malpass & Lindsay, 1999), and Binomial Probability (Tredoux, 1999). Mansour et al. (2017) provided a summary of these mock-witness measures and how to calculate them in their Supplementary Table 2.

Mock-witness measures can be regarded as content valid if their estimates vary in expected ways across different levels of similarity between a suspect and fillers in a lineup. That is, mock-witness measures would be content valid if, as overall suspect-filler similarity increases, lineup size increases whereas lineup bias decreases. Considering the direction yielded by the formula for each of the mock-witness measures, lineups whose fillers more closely resemble a suspect should be associated with higher estimates of Effective Size, Tredoux's *E*, ALM, Functional Size, and Binomial Probability and with lower estimates of Proportion of Suspect Selections, Suspect Bias, and Defendant Bias.

Corey et al. (1999) assessed the content validity of some of these mock-witness measures by manipulating the extent to which a suspect stood out from fillers in their lineups and

examining whether estimates of mock-witness measures differed across levels of their manipulation. Their lineups consisted of six lineup members including the guilty suspect. The suspect's photograph depicted him squinting (unlike the other fillers; biased) or not squinting (like other fillers; unbiased). Their results partially supported the content validity of mock-witness measures: In the unbiased lineup (cf. the biased lineup), mock witnesses were more likely to choose the fillers, which increased estimates of Functional Size, Tredoux's *E*, and ALM and decreased estimates of Suspect Bias. However, changes in the contents of the descriptions affected the relationship between the mock-witness measures and manipulated lineup fairness. As the descriptions of the suspect became less precise, Tredoux's *E* and ALM no longer varied with the level of lineup bias.

Brigham, Ready, and Spier (1990) also manipulated whether their lineups were biased versus unbiased and considered how this manipulation influenced Functional Size and Effective Size. They had undergraduate participants rate 80 photos of men for their similarity to each target photo. The authors randomly selected fillers from the middle third of the similarity-score distribution for the biased lineups, whereas a police officer who frequently constructed lineups selected high-similarity fillers (based on the similarity scores) for the unbiased lineups. Consistent with the expectation that mock-witness measures are content valid, the unbiased lineups yielded higher estimates of Functional Size and Effective Size than did the biased lineups.

McQuiston and Malpass (2002) also tested the content validity of mock-witness measures, but with sequential lineups when lineup fairness, lineup instructions, and criterion instructions were manipulated. For the lineup-fairness manipulation, fillers who matched the suspect's description completely were included in a fair lineup while fillers who matched

approximately half of the description were included in an unfair lineup. They concluded that the mock-witness paradigm may be valid with sequential lineups under restricted circumstances, such as when mock witnesses only make one lineup choice. They found that Proportion of Suspect Selections was higher in the unfair lineup than the fair lineup, and that methodological factors, such as instruction bias (a biased instruction forcing mock witnesses to choose a lineup member versus an unbiased instruction giving a not there option) and decision criterion (a high criterion instruction emphasizing the importance of correct identifications versus a low criterion instruction mentioning that this is not a real case and there will be no consequences of participants' identifications), did not affect the distribution of mock-witness lineup choices.

Some researchers did not manipulate lineup fairness explicitly, but have demonstrated that estimates of mock-witness measures vary with lineup characteristics that could affect lineup fairness. For example, Proportion of Suspect Selections increased when one of the fillers had clearly visible facial scars (Buckhout, Rabinowitz, Alfonso, Kanellis, & Anderson, 1988) and when a suspect was placed between the two fillers who looked the least like the suspect from amongst the five fillers in a lineup (Gonzalez, Davis, & Ellsworth, 1995). In sum, researchers have examined content validity in a variety of ways and overall, the data seem to indicate that mock-witness measures are content valid.

Concurrent and predictive validity

Concurrent validity indicates a relationship between the measure in question and other methods for assessing the same domain while predictive validity indicates the measure in question has a significant relationship with specific future events or outcomes. Very little research has considered these types of validity in relation to mock-witness measures—we are aware of only one study each relevant to these types of validity. If mock-witness measures have

concurrent validity, lineup fairness estimates for a lineup using mock-witness measures will predict other estimates of lineup fairness (e.g., human ratings or software algorithms). Brigham and Brandt (1992) compared lineup fairness derived from the mock-witness paradigm with lineup fairness based on more direct evaluations made by law officers and undergraduates. The authors had law officers and undergraduates view 23 lineups and rate the overall fairness of each of the lineups on a 6-point scale. Estimates of a mock-witness measure, which was a composite of Proportion of Suspect Selections, Effective Size, and ALM, predicted the law officers and undergraduates' fairness ratings, $r(21) = .42$.

Mock-witness measures can be regarded as predictively valid if they predict the selection behaviors of eyewitnesses viewing the lineup. Lindsay, Smith, and Pryke (1999) examined whether mock-witness measures predicted eyewitness identification performance, and found that lineup bias measures (e.g., Proportion of Suspect Selections and Defendant Bias), but not lineup size measures (e.g., Effective Size, and ALM), significantly predicted false identification rates in target-absent (TA) lineups. Tredoux (1999) proposed that the failure of lineup size measures to predict false identification rates might be due to the research design used by the authors—Lindsay et al. did not designate an innocent suspect in their TA lineups, but treated each lineup member as an elected suspect, which led to the computation of identical Effective Size for each lineup, making it impossible to predict performance For target-present (TP) lineups, neither the lineup bias measures nor the lineup size measures significantly predicted correct identification rates.

Convergent and discriminant validity

The logic behind convergent and discriminant validity is that different measures of the same domain should correlate with each other while measures of conceptually distinct domains

should not correlate with each other (Campbell & Fiske, 1959). Mansour et al. (2017) investigated the convergent and discriminant validity of mock-witness measures using a mock-witness task. As mentioned earlier, they considered lineup size and lineup bias as separate dimensions of lineup fairness with Effective Size, Tredoux's *E*, and ALM as measures of lineup size, and Proportion of Suspect Selections, Functional Size, Suspect Bias, Defendant Bias, and Binomial Probability as measures of lineup bias. They found significant intradimensional correlations within lineup size measures and within some lineup bias measures, supporting the convergent validity of mock-witness measures. The discriminant validity of the mock-witness measures was also partially supported by small or nonsignificant interdimensional correlations across the lineup size and lineup bias measures. Brigham, Meissner, and Wasserman (1999) also examined the convergent and discriminant validity of lineup size measures (Effective Size and ALM) and lineup bias measures (Proportion of Suspect Selections, Functional Size, and Suspect Bias)—which they measured across lineups employed in 18 criminal cases. Consistent with Mansour et al.'s (2017) intradimensional correlations, the correlation between Effective Size and ALM was large ($r = .80$, $p < .001$). The lineup size measures also produced a high degree of agreement (80-90%) in categorical classifications of lineups into fair or unfair lineups, whereas the agreement was much lower for lineup bias measures (33-50%). However, unlike Mansour et al.'s interdimensional correlations, lineup size measures significantly correlated with a lineup bias measure. Effective Size and ALM strongly correlated with Proportion of Suspect Selections (rs = -.77, and -.76, respectively, ps < .001).

The Current Study

The studies reviewed above generally tested validity using laboratory experiments—for example, Corey et al. (1999) manipulated the extent to which a suspect stands out from fillers in a lineup and examined whether mock-witness measures vary in expected ways with the manipulated lineup bias. This experimental approach is proper to investigate the relationship between mock-witness measures and eyewitness performance systematically. However, most of those experimental studies used a small number of lineups as the basis for their analyses, which resulted in low variability in the individual faces used for each of the studies.

Therefore, in the current study, we adapted a meta-analytic approach to advance the small literature on the validity of mock-witness measures for assessing lineup fairness. To this end, we built a database of eyewitness studies that estimated lineup fairness using mock-witness measures and that provided eyewitness performance associated with the lineups. The database allowed us to include a large set of lineups and test the validity of mock-witness measures from a meta-analytic perspective. Using this approach permitted us to focus on three types of validity—predictive, convergent, and discriminant. We were unable to investigate concurrent validity because researchers generally use the mock-witness paradigm only to measure lineup fairness within a study, not using other methods (e.g., human ratings or software algorithm) together. Therefore, they tend to provide mock-witness measures without lineup fairness indices produced by different methods, which are necessary for an analysis of concurrent validity.

To test predictive validity, we examined whether estimates of lineup fairness measured by the mock-witness paradigm can reliably predict eyewitness behaviors (e.g., suspect and filler choices) associated with lineups. Unlike Mansour et al. (2017) and Brigham et al. (1999) who investigated convergent and discriminant validity within subdimensions of mock-witness

measures (i.e., within lineup size and lineup bias), we investigated the validity of mock-witness measures in two domains which influence eyewitness performance—lineup fairness and the strength of eyewitness' memory for the perpetrator (memory strength). Given that eyewitness performance derives from a combination of the memory strength for the perpetrator and contextual biases (e.g., lineup fairness, influences of the lineup administrators, and lineup instructions) and that memory strength and contextual biases are conceptually independent, mock-witness measures should correlate with other indices of lineup fairness but not with indices reflecting memory strength. Therefore, the present study tested the convergent and discriminant validity of mock-witness measures by examining the intradomain correlations of mock-witness measures with other indices measuring lineup fairness (convergent validity) and interdomain correlations of mock-witness measures with indices reflecting memory strength (discriminant validity).

Method

Database Construction

We searched PsycINFO, Google Scholar, EBSCO, and our internal database for empirical eyewitness studies that used the mock-witness paradigm to measure lineup fairness. Keywords for the searches included mock-witness, Effective Size, Functional Size, Tredoux's *E*, lineup fairness, and lineup bias. We found 227 eyewitness studies relevant to at least one of the keywords.

Studies were included in the database if they (a) measured the fairness of lineups using the mock-witness paradigm; (b) used at least one of Effective Size, Functional Size, and Tredoux's *E*; and (c) provided information about eyewitness performance associated with the lineups, such as suspect identification (ID) rates, filler ID rates, and rejection rates. Of the 61

studies that met the inclusion criteria, we excluded 18 that did not calculate mock-witness measures separately for TP and TA lineups—the excluded studies provided the mean or the range of estimates of mock-witness measures of TP and TA lineups. However, when a study provided at least one of the mock-witness measures in either TP or TA lineups, we included it in our database. Therefore, 43 studies were included in our dataset (see Appendix).

Coding

We coded three types of variables of interest in this meta-analysis: mock-witness measures, eyewitness performance, and influences on eyewitness performance (specifically, memory strength and lineup fairness). Two coders generated separate codings of the mock-witness measures; the coding agreement rate between the two coders was .90. We used Lee's (2019) meta-analysis database to code eyewitness performance as well as study characteristics reflecting memory strength and lineup fairness—exposure time and filler similarity. In the meta-analysis database, the coding agreement rate was .81 for dependent variables including response rates, and .91 for study characteristics including exposure time and filler similarity.

Mock-witness measures. We coded Tredoux's *E* (Tredoux, 1998), Effective Size (Malpass, 1981), and Functional Size (Wells et al., 1979) in TP and TA lineups as mock-witness measures. Although there are other types of lineup fairness indices, these are the most frequently reported and other measures were not used frequently enough to allow for meaningful analysis. We looked at a boxplot for each measure and identified one extreme outlier, which was outside of the range between [the 3rd quartile + 3 × interquartile range] and [the 1st quartile – 3 × interquartile range], for Functional Size in TP lineups (Functional Size = 20.20; Brigham et al., 1982). This extreme value was removed from our analyses.

To increase our power, the standardized z-scores of Tredoux's *E* and Effective Size were combined into one index, Lineup Size—in general, researchers rarely provide values of Tredoux's *E* and Effective Size together in a study (only Beresford & Blades (2006) provided estimates of both for the same lineup). Mansour et al. (2017) and Malpass, Tredoux, and McQuiston-Surrett (2007) demonstrated that Tredoux's *E* and Effective Size measure a single, related construct by providing evidence that they are strongly and significantly correlated with one another ($.99 \geq r\text{s} \geq .97$). Therefore, we computed *M* and *SD* for each of Tredoux's *E* and Effective Size, computed z-scores for each of Tredoux's *E* and Effective Size, and combined them into one variable because the computed z-scores were standardized. For Beresford and Blades' study, the standardized Tredoux's *E* was used. Therefore, further analyses were conducted with two types of mock-witness measures—Lineup Size, which is a measure of lineup size, and Functional Size, which is a measure of lineup bias.

Eyewitness performance. Eyewitness performance variables included overall response rates, such as suspect ID rates, filler ID rates, and rejection rates in TP and TA lineups. We also coded multiple measures of overall discriminability, following the multi-d′ model (Lee & Penrod, 2019). The multi-d′ model proposes five types of eyewitness discriminability in lineups—d′(GI), the discriminability of a guilty suspect from an innocent suspect; d′(GF_p), the discriminability of a guilty suspect from fillers; d′(IF_a), the discriminability of an innocent suspect from fillers; d′(F_aF_p), the differential appeal of fillers between TP and TA lineups; and d′(R_aR_p), the discriminability between the presence and absence of the perpetrator. Each of the d′ measures is calculated using the difference between the z-transformed proportions associated with the two components. For example, d′(GF_p) and d′(IF_a) are calculated as the z-transformed guilty suspect ID rate minus the z-transformed TP filler ID rate, and the z-transformed innocent

suspect ID rate minus the z-transformed TA filler ID rate, respectively. In addition, we coded response rates and discriminability values across conditions of each independent variable.

Influences on eyewitness performance (memory strength and lineup fairness). To test the convergent and discriminant validity of the mock-witness measures, indices of memory strength for the perpetrator and lineup fairness were generated from eyewitness performance and study characteristic variables. The current study assessed convergent and discriminant validity of the mock-witness measures by examining whether mock-witness measures correlate with other indices measuring lineup fairness, but not with indices measuring memory strength for the perpetrator. For these analyses, three families of variables that reflect memory strength and lineup fairness were created. The first two families were calculated from the eyewitness performance variables (i.e., raw response rates) and the third family was coded based on study characteristics in each study. Table 1 summarizes the calculations and variable representations.

The multi-d′ model. The first family of the variables was calculated based on the multi-d′ model (Lee & Penrod, 2019). As mentioned earlier, **$\mathbf{d'(R_aR_p)}$** estimates the ability to detect the perpetrator in a lineup, and therefore can be regarded as a proxy measure of memory strength for the perpetrator. $d'(R_aR_p)$ is computed by taking the z-transformed TA rejection rate minus the z-transformed TP rejection rate. Although **$\mathbf{d'(IF_a)}$** indicates the discriminability of an innocent suspect from TA fillers, it can also be regarded as TA lineup bias because the discriminability simply comprises the perceived similarity between an innocent suspect and fillers without the influence of the memory strength. However, given that $d'(GF_p)$ is the discriminability of the perpetrator from fillers, this discriminability derives from two components—memory strength for the perpetrator and the perceived similarity between the perpetrator and fillers (i.e., TP lineup bias). That is, $d'(GF_p)$ increases when memory for the perpetrator is strong and/or when the

similarity between the perpetrator and fillers is weak. To separate out memory strength from $d'(GF_p)$, the multi-d′ model subtracts $d'(R_aR_p)$ from $d'(GF_p)$, and proposes that the parameter, **$d'(GF_p) - d'(R_aR_p)$** reflects TP lineup bias. Therefore, in the first variable family, $d'(R_aR_p)$ was used as an index measuring the memory strength for the perpetrator and $d'(GF_p) - d'(R_aR_p)$ and $d'(IF_a)$ were used as indices of lineup fairness for each of the TP and TA lineups respectively.

Strategy-based breakdown. The second family of variables reflecting memory strength for the perpetrator and lineup fairness are calculated using a strategy-based breakdown of eyewitness types—reliable eyewitnesses and guessers (Penrod, 2003). Calculation of these variables is coupled with a simplifying assumption—because eyewitness researchers randomly assign participants to either TP or TA lineups, we can use performance in one condition to make inferences about performance in the other condition. The analysis classifies participants into reliable eyewitnesses and guessers based on whether they made decisions indicative of a strong memory for the perpetrator. Reliable eyewitnesses are eyewitnesses who would make correct responses whether randomly assigned to TP or TA lineups (i.e., guilty suspect IDs in TP lineups and rejections of TA lineups) based on their strong memory for the perpetrator. Eyewitnesses other than reliable eyewitnesses are regarded as guessers who make their decision from a weak memory. Notably, some guessers make correct responses (i.e., lucky guessers).

In this view, eyewitnesses who reject TP lineups are guessers while eyewitnesses who reject TA lineups consist of both reliable eyewitnesses and guessers. We can expect that guessers who rejected TP lineups would similarly reject TA lineups because they rejected lineups which included the perpetrator who is nearly always replaced by someone who only bears a resemblance to the perpetrator. Therefore, we can subtract the rejection rate for TP lineups (i.e., guessers) from the rejection rate in TA lineups (i.e., reliable eyewitnesses + guessers) to obtain

the proportion of eyewitnesses with a strong memory for the perpetrator (i.e., **Reliable Eyewitnesses**). For example, in Table 2, Reliable Eyewitnesses is computed by .55 – .30 = .25.

According to the strategy-based breakdown, lineup bias in TP lineups (**TP bias**) is the ratio of guessers' guilty suspect ID rates to the average ID rate (the average ID rate = the total ID rate / nominal lineup size) in TP lineups. Given that the guilty suspect identifiers consist of reliable eyewitnesses and guessers, guessers' guilty suspect ID rates can be calculated by subtracting the proportion of reliable eyewitnesses (.25 as calculated above) from the total guilty suspect ID rates (i.e., .50 in Table 1). Because all positive identifiers in TA lineups are guessers, lineup bias in TA lineups (**TA bias**) is the ratio of innocent suspect ID rates to the average ID rate in TA lineups. TP bias for the rates provided in Table 2 is computed by (.50 – .25)/(.70/6) = 2.14 and TA bias is computed by .10/(.45/6) = 1.33.

Study characteristics. The last family of variables reflecting memory strength and lineup fairness were coded based on study characteristics. **Exposure Time** of the perpetrator at the encoding phase (e.g., the length of a mock-crime video) was used as a proxy variable reflecting memory strength for the perpetrator—longer exposure time is associated with a stronger memory for the perpetrator. Exposure time was coded in seconds. Although there were other study characteristics that could affect the memory strength, such as whether the perpetrator's face was disguised during the crime or whether the perpetrator had a distinctive facial feature (e.g., mole or tattoo), these study characteristics varied little across the studies included in the database—most of the included studies did not disguise perpetrators' faces and did not include perpetrators with distinctive features. When a study manipulated exposure time with other independent variables, the exposure time for datapoints of the other independent variables was coded with the average exposure time.

The similarity between a suspect and fillers (**Filler Similarity**) was included as a study characteristic reflecting lineup fairness. This variable was coded as high (= 1), moderate (= 0), or low (= -1). When a study manipulated filler similarity as an independent variable, this variable was coded as high, moderate, or low, based on the authors' operationalization. When filler similarity was manipulated with other independent variables, filler similarity was coded as moderate for the datapoints of the other independent variables. In the absence of the relevant information, filler similarity was coded as moderate.

Results

Predictive Validity

To test the predictive validity of mock-witness measures, we investigated whether mock-witness measures reliably predict eyewitness performance. We expected that, if mock-witness measures have predictive validity, fairer lineups estimated by the mock-witness measures would yield lower suspect ID rates, higher filler ID rates, and lower discriminability of a suspect from fillers (i.e., $d'(GF_p)$ and $d'(IF_a)$). Therefore, we examined correlations among mock-witness measures, eyewitnesses' response rates and discriminability of a suspect from fillers.

As shown in Figure 1, higher Lineup Size was associated with higher Filler IDs ($r(88) = .32$ $p = .002$) and lower discriminability of a suspect from fillers ($r(88) = -.31$, $p = .003$); higher Functional Size was also associated with higher Filler IDs ($r(19) = .45$, $p = .05$) and lower discriminability of a suspect from fillers ($r(19) = -.49$, $p = .03$). However, the mock-witness measures did not significantly correlate with Suspect IDs, contrary to our prediction.

Additionally, the mock-witness measures did not correlate with rejections ($r(88) = -.13$, $p = .22$ for Lineup Size; $r(19) = .06$, $p = .80$ for Functional Size). The nonsignificant correlations between Rejections and the estimated lineup fairness are consistent with findings from Fitzgerald

et al.'s (2013) meta-analysis. Fitzgerald et al. (2013) demonstrated that, unlike suspect and filler ID rates, rejection rates did not significantly vary across different levels of the actual lineup fairness (for more details, see their Table 1).

Overall, Lineup Size and Functional Size demonstrated a similar pattern of results. The two mock-witness measures had a modest ability to predict Filler IDs and the discriminability of a suspect from fillers, but not Suspect IDs and Rejections.

Moderation of Target Presence. In order to investigate whether the ability of mock-witness measures to predict eyewitness performance differs between TP and TA lineups, we conducted regression analyses—we regressed each DV of eyewitness performance on Lineup Size, Target Presence, and their interaction-effect term (see Table 3). Because of the small number of observations for Functional Size, we conducted the moderation analysis with Lineup Size only. However, we provide scatterplots and correlations using both Lineup Size and Functional Size in Supplemental Material 1 to give a comprehensive view of the moderating effects of Target Presence. The interaction effect of Target Presence × Lineup Size was statistically significant for Filler IDs ($p = .03$) and marginally significant for Discriminability ($p = .06$). Simple slope tests (Aiken & West, 1991) demonstrated that Lineup Size was a significant predictor of Filler IDs and Discriminability for TA lineups ($B = .10$, $p < .001$ for Filler IDs; $B = -.56$, $p = .001$ for Discriminability) but not for TP lineups ($B = .03$, $p = .19$ for Filler IDs; $B = -.18$, $p = .13$ for Discriminability). Although the interaction of Target Presence × Lineup Size was not significant for Suspect IDs and Rejections, the same pattern of simple slopes was found for those DVs—higher Lineup Size was associated with lower Suspect IDs or Rejections only for TA lineups but not for TP lineups (see Table 3).

Moderation of Memory Strength in TP lineups. The inability of Lineup Size to predict TP lineup decisions in the regression analyses led us to consider potential reasons why TP and TA lineups would differ in terms of lineup fairness. We reasoned that the presence of a perpetrator in lineups may weaken the ability of mock-witness measures to predict eyewitness behaviors. We hypothesized that this may occur because memory strength for the perpetrator moderates the relationship. Specifically, we expected that for TP lineups, mock-witness measures would predict eyewitness performance when memory strength is moderate but not weak or strong. The logic of this prediction is as follows. When eyewitnesses have a weak memory for the perpetrator, their choices will be distributed across all lineup members regardless of lineup fairness because the similarity between the perpetrator and fillers cannot be a contextual cue for eyewitnesses who do not remember the perpetrator's appearance. In the same vein, when eyewitnesses have a strong memory for the perpetrator, their choices will concentrate on the perpetrator in a lineup, regardless of lineup fairness, because they will rely on their memory and have no need to rely on other cues, including the similarity between the perpetrator and fillers. This prediction is consistent with prior studies suggesting that eyewitnesses with a strong memory for the perpetrator are rarely influenced by contextual cues, such as lineup instructions (e.g., appearance-change instructions; Charman & Wells, 2007) and lineup presentation modes (e.g., simultaneous versus sequential lineups; Lindsay, Mansour, Beaudry, Leach, & Bertrand, 2009; Penrod, 2006).

Therefore, to test whether memory strength moderates the ability of mock-witness measures to predict eyewitness performance in TP lineups, we conducted regression analyses to test the 3-way interaction of mock-witness measures, target presence, and memory strength on eyewitness performance. We used $d'(R_aR_p)$ as a proxy measure of memory strength, and cases in

the current database were categorized into three groups (weak, moderate, or strong memory) based on the values of $d'(R_aR_p)$. The cut-off values of $d'(R_aR_p)$ for the categorization were the first and third quartiles of $d'(R_aR_p)$ in Lee (2019). We did not use quartile values of $d'(R_aR_p)$ from the current database as Lee's database provides a more comprehensive and normal distribution because of its large number of observations (see Figure 2). Thus, we categorized memory strength into three groups based on the first and third quartiles of $d'(R_aR_p)$ in Lee (2019): $d'(R_aR_p) < 0.36$ for weak memory, $0.36 \leq d'(R_aR_p) < 0.99$ for moderate memory, and $d'(R_aR_p) \geq 0.99$ for strong memory.

Unlike the moderation analysis with target presence above using the overall eyewitness performance values only, this moderation analysis used eyewitness performance values for each level of a study's independent variables as well as the overall values[1]. This strategy widened the range of $d'(R_aR_p)$ and increased the total number of observations. For example, Carlson, Young, Weatherford, Carlson, Bednarz, and Jones (2016) manipulated the exposure time of a target face as 3 seconds versus 10 seconds. If we include their overall eyewitness performance values only in the current analysis, $d'(R_aR_p)$ of Carlson et al.'s (2016) study will have one datapoint ($d'(R_aR_p) = .97$), which is computed from the overall rejection rates in TP and TA lineups. However, when also including $d'(R_aR_p)$ computed from TP and TA rejection rates for each level of the exposure-time variable, $d'(R_aR_p)$ has three datapoints and the value range of the index increases (i.e., overall $d'(R_aR_p) = .97$; $d'(R_aR_p) = .78$ for the 3s condition; $d'(R_aR_p) = 1.04$ for the 10s condition).

We regressed each of the eyewitness performance DVs on Lineup Size, Target Presence(N for TP lineups = 64, N for TA lineups = 31), and memory strength, and their 2-way

[1] Note that there is a dependency in these analyses to the extent that the lineup fairness measures are sometimes based on responses to the same mock crime/lineup materials. This can be true both when multiple measures are obtained from a single study or publication and when researchers use the same materials across studies or publications.

and 3-way interaction terms. However, we did not find a significant 3-way interaction. Results of the regression analyses are provided in Supplemental Material 2. Because of the small number of observations for Functional Size, we did not conduct the same regression analysis with Functional Size.

Although the expected 3-way interaction effect was not significant, we looked at correlations between the mock-witness measures and eyewitness performance for each of TP and TA lineups across different memory-strength levels to determine whether the pattern of correlations was consistent with our expectation. The full correlations and scatter plots are available in Figure 3a to 3c (because the scatter plots are based on small sample sizes, theses plots should be interpreted cautiously). For TP lineups, Lineup Size correlated significantly with Filler IDs in the weak and moderate memory conditions ($r(44) = .33$, and $r(46) = .29$ respectively, ps $\leq .05$). Lineup Size also had significant correlations with Suspect IDs, $r(47) = -.34$, $p = .02$, and $d'(GF_p)$, $r(46) = -.39$, $p < .01$, only in the moderate memory condition. Functional Size correlated with Filler IDs in the weak and moderate memory condition ($r(13) = .54$, $p = .06$ and $r(16) = .73$, $p = .001$ respectively) and $d'(GF_p)$ only in the moderate memory condition ($r(16) = -.84$, $p < .001$).

For TA lineups, Lineup Size had moderate to strong correlations with Filler IDs and $d'(IF_a)$ for all the three memory conditions. Notably, as the memory for the perpetrator became stronger, the correlations also became stronger—$r_{\text{Filler ID}}(30) = .58$, $r_{\text{Discriminability}}(30) = -.45$ for the weak memory group; $r_{\text{Filler ID}}(27) = .78$, $r_{\text{Discriminability}}(27) = -.83$ for the moderate memory group; $r_{\text{Filler ID}}(4) > .99$, $r_{\text{Discriminability}}(4) < -.99$ for the strong memory group, all ps $\leq .01$—however, because of the small number of cases in the strong memory group, these results must

be cautiously interpreted. Lineup Size had a significant correlation with Suspect IDs only for the moderate memory condition ($r(28) = -.77$, $p < .001$).

In sum, although our regression analyses failed to demonstrate a significant 3-way interaction of mock-witness measures, target presence, and memory strength on eyewitness performance, correlation analyses revealed potential interactions. The analyses for Functional Size are based on a quite small samples, however, and should be interpreted cautiously. For TP lineups, mock-witness measures consistently had significant correlations with eyewitness performance when memory was moderate, but sometimes also when memory was weak. For TA lineups, mock-witness measures generally had moderate to strong correlations with eyewitness behaviors (Filler IDs and $d'(IF_a)$) regardless of the memory strength for the perpetrator. Broadly, the results are consistent with our prediction that memory quality moderates the relationship between mock-witness measures and lineup performance in TP lineups.

Convergent and Discriminant Validity

To test the convergent and discriminant validity of mock-witness measures, we used a multitrait-multimethod (MTMM) correlation matrix (Campbell & Fiske, 1959). A MTMM correlation matrix includes similar traits measured by different methods (i.e., monotrait-heteromethod unit) to assess convergent validity and different traits measured by the same method (i.e., heterotrait-monomethod unit) and different methods (i.e., heterotrait-heteromethod unit) to assess discriminant validity. The logic behind the MTMM correlation matrix is that, when testing the construct validity of a measure, the measure should correlate with similar traits measured by other methods (convergent validity) and not correlate with different traits measured by the same or different methods (discriminant validity).

As mentioned earlier, eyewitness performance on lineups is affected by memory strength and contextual biases (e.g., lineup fairness), and these components are conceptually independent. Thus, we created the three families of variables (i.e., the multi-d′ model, strategy-based breakdown, and study characteristics) reflecting memory strength and lineup fairness. Given that mock-witness measures are intended to estimate lineup fairness, they should correlate with the other lineup fairness measures but not memory strength measures.

For the MTMM correlation analyses in Table 4, all correlation coefficients, except those involving Filler Similarity, were Pearson correlation coefficients. Given that Filler Similarity was an ordinal variable (i.e., high = 1, moderate = 0, and weak = -1), we computed Spearman's correlation coefficients to evaluate the relationship between Filler Similarity and other measures.

The MTMM correlation matrix (see Table 4) supported the convergent and discriminant validity of the mock-witness measures. The mock-witness measures significantly correlated with indices measuring lineup fairness ($-.75 \leq rs \leq -.26$, all $ps < .05$), but not with indices measuring the memory strength for the perpetrator ($-.20 \leq rs \leq .07$, all $ps > .09$).

Notably, compared to Lineup Size, Functional Size had stronger correlations with the lineup fairness measures of the multi-d′ model ($r(66) = -.45$ for Lineup Size and $r(10) = -.75$ for Functional Size) and Penrod's (2003) strategy-based breakdown ($r(66) = -.26$ for Lineup Size and $r(10) = -.65$ for Functional Size), although the difference was not statistically significant ($z = -1.23$, $p = .11$ for the multi-d' model; $z = -0.37$, $p = .36$ for the strategy-based breakdown). The stronger correlations with Functional Size might occur because the measures from the multi-d′ model and the strategy-based breakdown are more conceptually related to the domain measured by Functional Size than Lineup Size. Considering the formulae for the lineup fairness measures from the multi-d′ model and strategy-based breakdown, their estimates can be regarded as

reflecting the extent of which a suspect stands out from other fillers in a lineup, rather than the number of lineup members who are sufficiently similar to the perpetrator to be sometimes chosen (we will discuss this in more detail in the discussion). Furthermore, the non-significant correlation between Lineup Size and Functional Size ($r(24) = .22$, $p = .31$) in Table 4 indicates that they measure different subdomains of lineup fairness. Given that the dimension measured by the lineup fairness measures of the multi-d′ model and the strategy-based breakdown are more closely associated with lineup bias than lineup size, the stronger correlations between Functional Size (cf. Lineup Size) and the lineup fairness measures from the multi-d' model and the strategy-based breakdown provide evidence for the distinct subdimensions of lineup fairness. However, considering the small sample of Functional Sizes, the results involving Functional Size should be cautiously interpreted.

In addition, the strong intradomain correlations between the multi-d′ model and strategy-based breakdown indicate that their indices may be used interchangeably ($r(57) > .99$ for the memory strength; $r(66) = .93$ for the lineup fairness).

Discussion

The current study investigated the validity of mock-witness measures for assessing lineup fairness in eyewitness identification research, focusing on their predictive, convergent, and discriminant validity.

First, mock-witness measures predicted eyewitness performance significantly in TA but not TP lineups. Thus, mock-witness measures evidence predictive validity only for TA lineup decisions. This result is somewhat consistent with Lindsay et al. (1999) who found that TA but not TP lineup decisions were related to mock-witness measures for simultaneous lineups (more than 80% of our sample used simultaneous lineups). However, unlike their findings that filler ID

rates significantly correlated with lineup bias measures but not with lineup size measures in TA lineups, we found TA lineup decisions including filler ID rates significantly correlated with lineup size measures. The inconsistent results regarding the correlation between lineup size measures and filler ID rates might be because Lindsay et al.'s (1999) correlation analysis had lower power than our own. They used 15 TA lineups for the analysis, while we used 31 TA lineups.

We hypothesized that the reason for the mismatch between mock-witness measures and eyewitness performance in TP lineups may be due to a moderating effect of the eyewitnesses' memory strength for the perpetrator. That is, we expected mock-witness measures to predict eyewitness performance on TP lineups when memory strength was moderate but not weak or strong. Regardless of the actual similarity between the guilty suspect and fillers in a TP lineup, choices by eyewitnesses with a poor memory for the perpetrator can be expected to be distributed across all the lineup members because similarity is not a cue for eyewitnesses who do not remember the perpetrator's face. Eyewitnesses with a strong memory for the perpetrator should concentrate on the guilty suspect because they can be expected to rely on their memory rather than contextual cues (Charman & Wells, 2007; Lindsay et al., 2009; Penrod, 2006). Although our regression analyses failed to find a significant 3-way interaction of Lineup Size, target presence, and memory strength on eyewitness performance, correlation analyses gave us a hint about possible interaction effects involving memory strength. Lineup Size yielded significant correlations with all the eyewitness performance indices (i.e., filler and suspect IDs, and the discriminability of a guilty suspect from fillers) for TP lineups in the moderate memory condition. We also sometimes found correlations between mock-witness measures and eyewitness performance indices for TP lineups in the weak memory condition. These latter

unexpected correlations might have occurred because we used a fairly rough categorization for memory strength and the weak memory group might have possessed some memory for the perpetrator. It would be desirable to test the moderation effect of memory strength with a more precise operationalization of memory strength and greater power in future research.

Our analyses also considered the convergent and discriminant validity of mock-witness measures by examining a MTMM correlation matrix, including multiple indices relevant to eyewitness performance—lineup fairness and memory strength. Since there are no objective cutoff values of *r* coefficients to establish convergent and discriminant validity (DeVellis, 2012), researchers use arbitrary criteria to determine whether their measures have validity. For example, Post (2016) expected strong correlations (*r*s >. 60 or .70) and weak or not significant correlations (*r*s < .30 or .40) for convergent and discriminant validity, respectively, in a MTMM matrix. Brown (2006) regarded a correlation above .80 as indicating a lack of discriminant validity. Because of these arbitrary criteria, Engellant, Holland, and Piper (2016) considered relative size and significance of *r* coefficients between heterotrait-heteromethod, heterotrait-monomethod, and monotrait-heteromethod cells. They suggested that correlations in monotrait-heteromethod cells should be higher than those in heterotrait-monomethod or heterotrait-heteromethod cells for convergent validity, and correlations in heterotrait-heteromethod cells should be not significant (or barely reach significance) for discriminant validity. Our MTMM correlation matrix demonstrated that Lineup Size and Functional Size strongly correlate with indices of lineup fairness ($-.75 \leq r\text{s} \leq -.26$, all $p\text{s} < .05$) but not with indices of memory strength ($-.20 \leq r\text{s} \leq .07$, all $p\text{s} > .09$). Given Engellant et al.' (2016) suggested criteria, our results can be regarded as supporting the convergent and discriminant validity of the mock-witness measures.

In addition, the non-significant correlation between Lineup Size and Functional Size in the MTMM matrix provides evidence that they measure different subdimensions of lineup fairness—lineup size (the number of plausible members in the lineup) and lineup bias (the extent of which a suspect stands out from fillers in the lineup). Since Malpass (1981) first distinguished between these subdimensions, they have been in use, yet, limited data has supported the theoretical distinction. Brigham et al. (1999) reported a high correlation between lineup size and bias (though they used a different bias measure from us) but their analysis was based on suspect identifications and could not differentiate between TA and TP lineups because they were using archival data. Our results are more in line with Mansour et al. (2017) who found moderate to strong intradimensional correlations and weak to moderate interdimensional correlations. Although these dimensions are differentiable, they are clearly not independent.

Furthermore, compared to Lineup Size, Functional Size had stronger correlations with the lineup fairness measures of the multi-d′ model and strategy-based breakdown, which are more conceptually related to lineup bias than lineup size. The computation of $d'(GF_p) – d'(R_aR_p)$ and $d'(IF_a)$ is based on a difference in mock witness' choosing rates between a suspect and fillers.TP bias and TA bias in the strategy-based breakdown are computed as a ratio of guessers' suspect ID rate to the average ID rate in a lineup. Considering the computations focus on the imbalance between suspect and filler ID rates of guessers, the dimensions estimated by these measures are closer to the extent to which a suspect stands out from fillers than the number of plausible members in a lineup. Our findings suggest that researchers may use the lineup fairness measures of the multi-d′ model and the strategy-based breakdown for a post-assessment of lineup bias. Note that, unlike mock-witness measures, these measures can be directly computed from eyewitness performance. Furthermore, researchers may use the lineup fairness measures of the

multi-d' model with mock-witness measures to cross-check the appropriateness of their lineup stimuli. However, our data also indicate that one thing that these approaches may not do well is provide insight into lineup size, so there is certainly scope for further advancement.

Limitations and Future Directions

A point for consideration in the present study is that our database was small. In developing our database, we found a number of instances where researchers reported mock-witness measures in a way that did not allow us to include them in our analyses (e.g., they provided ranges rather than specific values). In addition, researchers may be using mock-witness measures but not always reporting them, particularly if they conclude from the measures that their lineups are not clearly fair. This "file-drawer" problem could obscure the true impact of lineup fairness such that our analysis only represents circumstances where lineups perform well according to the mock-witness measures.

The lack of data availability also limited our analyses. For example, we did not have sufficient data to test the predicted three-way interaction of lineup bias, target presence, and memory strength and the three-way interaction of lineup size, target presence, and memory strength was not significant—likely in large part due to the small number of observations for TA lineups. Nonetheless, our data do indicate that memory strength for the perpetrator moderated the ability of mock-witness measures to predict eyewitness performance in TP lineups, . We expect the similarity between a perpetrator and an innocent suspect to serve as a moderator in TA lineups—specifically, mock-witness measures should predict eyewitness behaviors significantly in TA lineups when the similarity is low or moderate, but not strong. The logic is as follows. When an innocent suspect closely resembles the perpetrator, the relationship between mock-witness measures and eyewitness behaviors in TA lineups would be non-significant because the

relationship is moderated by memory strength for the perpetrator, like the moderating effect of memory strength in TP lineups—the overall relationship between mock-witness measures and eyewitness performance in TP lineups was not significant, but the relationship became significant when the memory strength was restricted to moderate or weak. Despite the small number of observations, we conducted the moderation analysis of the similarity between a perpetrator and an innocent suspect for exploratory purposes, and the results were broadly consistent with our prediction. We provide results from the moderation analysis as a Supplemental Material 3 for readers who may be curious about the results. Considering the lack of data availability, we would encourage more researchers to calculate and report mock-witness measures and, like Mansour et al. (2017), encourage the reporting of the measures for both TA and TP lineups. We would also encourage researchers to report mock-witness measures for each lineup, rather than providing descriptive statistics such as means or a range.

Although our findings demonstrate that overall mock-witness measures can be regarded as having predictive, convergent, and discriminant validity for measuring the fairness of lineups, we cannot conclude that the mock-witness paradigm is an effective way to measure lineup fairness. That is, although mock-witness measures predicted TA lineup behaviors, memory strength for the perpetrator had to be considered in order to predict TP lineup behaviors; although the mock-witness measures correlated significantly with other indices of lineup fairness, these correlations varied considerably ($.19 \leq r\text{s} \leq .99$); and although the mock-witness measures correlated non-significantly with theoretically distinct measures (e.g., indices of memory strength), some of these correlations approached meaningful levels (e.g., functional size correlated at .22 with lineup size and lineup size correlated at -.18 with exposure time). Thus, we would encourage researchers to consider alternative ways of estimating lineup fairness.

One approach that has been reported in the literature is the use of a similarity rating of lineup members' appearance (e.g., Brewer & Wells, 2006), which is a potential method that could be substituted for the mock-witness paradigm. In the similarity rating method, people rate the similarity between the suspect and each of the fillers in a lineup using Likert-style scales. This method may be more reliable than the mock-witness paradigm because the raters view the perpetrator and then directly rate his or her similarity to other fillers.

Tredoux (2002) adapted principal-component analysis (PCA) to measure facial similarity. The PCA approach generates eigenfaces that represent images on standardized facial dimensions, which derive from a statistical analysis of a set of facial pictures. He demonstrated that facial similarity measured by the PCA approach strongly correlated with mock-witness measures (Tredoux's *E*, Effective Size, and Functional Size).

Another method would be to use scaling techniques, such as multidimensional scaling (MDS) models (Hirschberg, Jones, & Haggerty, 1978). MDS models suppose that human perceptions of the similarity between stimuli are based on a variety of latent features. For example, people perceive the similarity between human faces with multidimensional features, such as race, age, sex, face shape, eye size, and so on (Valentine, 1991). MDS models calculate similarity between faces using ratings of each of the multiple features. Human subjects or computer-based software programs, such as Betaface, are generally used to produce the ratings on the multidimensional features. Since facial recognition software programs rate each of the features with mathematical algorithms, their ratings may be more objective than human judgments. Bergold and Heaton (2018) used the software algorithm to measure lineup fairness, and proposed that facial recognition software may be a useful method for measuring lineup fairness because it is more objective than traditional methods.

Finally, a behavioral measure of choosing could be considered. For example, researchers could run a pilot test where people get a view of the target comparable to the view of real participants, then show them alternative lineups or a series of faces to gauge how often candidates for fillers are chosen. A similar approach might provide a group of participants with a very poor view of the target such that there would be no expectation that they would have a sufficient memory to identify the target, if present (Tredoux, 2019). The distribution of lineup choices could then be evaluated using mock-witness measures.

Considering these alternative methods for measuring lineup fairness, it would be desirable for future research to test the validity of alternatives and to investigate which method is the most valid approach for determining lineup fairness. In addition, the present study was unable to test the concurrent validity of mock-witness measures because researchers tend to use only a single approach for examining lineup fairness (i.e., the mock-witness task). Therefore, we also encourage others to test the concurrent validity of mock-witness measures with the alternative measures.

Implications

The validity of lineup fairness measures is important in that researchers often use the paradigm to check that their lineup stimuli are appropriate for their purpose, and then use these lineup stimuli to investigate the effects of various factors on eyewitness performance. Given that findings from academic research sometimes lead to changes in the policies of law enforcement agencies (e.g., double-blind lineups), the validity of lineup fairness measures should be investigated. If the measures used do not estimate lineup fairness in a valid way, researchers may draw conclusions about the impact of manipulated variables that are erroneous because of the nature of the lineups used, which may lead practitioners to implement inappropriate policies.

Furthermore, eyewitness experts are often asked to testify in court when the fairness of a lineup is questioned. A clear operationalization of lineup fairness, reliable ways of measuring it, and an appreciation of the factors that moderate it will allow eyewitness experts to better express the circumstances under which a lineup is fair and unfair. This testimony can then assist judges and jurors in judging the reliability of the identification evidence. Similarly, such information will help them sort through the literature when asked to testify about whether the circumstances of an identification are relevant. When researchers have found a specific factor to influence identification accuracy, and the lineups are considered fair using valid measures, experts can make stronger statements about the role that factor may have played in the current case.

Valid lineup fairness measures are also relevant at the level of lineup construction by police officers and prosecutors. Police departments are generally motivated to produce fair lineups because obviously unfair lineups (e.g., where the suspect's race is obviously different from the fillers) are more and more being considered as unacceptable evidence. Prosecutors likewise do not wish their cases to be thrown out of court. However, both are also concerned about making identifications overly challenging for eyewitnesses. An appreciation of what lineup fairness means and the factors that influence it could help police departments achieve this goal. For example, knowing that an eyewitness with a moderate memory is more strongly influenced by lineup size may lead police officers to take more care in ensuring that the lineup contains five (or however many) good fillers.

Ultimately, it would be useful for researchers to find a way to help other researchers, police officers, prosecutors, and the triers of fact assess what is an optimal level of fairness. The next step would be developing an easy-to-use tool that could be used at the lineup construction

stage or when evaluating the lineup in court. Our research is a first step towards this. Before we can determine what is an optimal level of fairness, we must ensure we can measure fairness.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage Publications, Inc.

*Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36-40. https://doi.org/10.1016/j.paid.2013.12.011

*Beresford, J., & Blades, M. (2006). Children's identification of faces from lineups: the effects of lineup presentation and instructions on accuracy. *The Journal of Applied Psychology*, *91*, 1102-1113. https://doi.org/10.1037/0021-9010.91.5.1102

Bergold, A. N., & Heaton, P. (2018). Does filler database size influence identification accuracy? *Law and Human Behavior, 42(3)*, 227-243. https://doi.org/10.1037/lhb0000289

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(1)*, 11-30. https://doi.org/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(5)*, 475-489. https://doi.org/10.1007/BF01044619

*Brigham, J. C., Maass, A., Larry, D. S., & Spaulding, K. (1982). Accuracy of eyewitness identifications in a field study. *Journal of Personality and Social Psychology, 42(4)*, 673-681. https://doi.org/10.1037/0022-3514.42.4.673.

Brigham, J. C., Meissner, C. A., & Wasserman, A. W. (1999). Applied issues in the construction

and expert assessment of photo lineups. *Applied Cognitive Psychology, 13(S1)*, S73-S92. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S73::AID-ACP631>3.3.CO;2-W

Brigham, J. C., Ready, D. J., & Spier, S. A. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology, 11(2)*, 149-163. https://doi.org/10.1207/s15324834basp1102_3

*Brigham, J. C., Verst, M. V., & Bothwell, R. K. (1986). Accuracy of Children's Eyewitness Identifications in a Field Setting. *Basic and Applied Social Psychology*, *7(4)*, 295-306. https://doi.org/10.1207/s15324834basp0704_4

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (1st ed.), New York, NY: Guilford Publications.

Bruer, K. C., Fitzgerald, R. J., Therrien, N. M., & Price, H. L. (2015). Line-up member similarity influences the effectiveness of a salient rejection option for eyewitnesses. *Psychiatry, Psychology and Law*, *22(1)*, 124-133. https://doi.org/10.1080/13218719.2014.919688

Buckhout, R., Rabinowitz, M., Alfonso, V., Kanellis, D., & Anderson, J. (1988). Empirical assessment of lineups: Getting down to cases. *Law and Human Behavior, 12(3)*, 323-331. https://doi.org/10.1007/BF01044388

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56(2)*, 81-105. https://doi.org/10.1108/02634500510597283

*Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *3(2)*, 45-53. https://doi.org/10.1016/j.jarmac.2014.03.004

*Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An Investigation of the Weapon Focus Effect and the Confidence–Accuracy Relationship for Eyewitness Identification. *Journal of Applied Research in Memory and Cognition*, *6(1)*, 82-92. https://doi.org/10.1016/j.jarmac.2016.04.001

*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(2)*, 118-128. https://doi.org/10.1037/1076-898X.14.2.118

Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamyeir, R. F., Clarson, M. A., & Wooten, A. R. (2019). Lineup fairness: propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive Research: Principles and Implications, 4(1)*, https://doi.org/10.1186/s41235-019-0172-5.

*Carlson, C. A., Young, D. F., Weatherford, D. R., Carlson, M. A., Bednarz, J. E., & Jones, A. R. (2016). The Influence of Perpetrator Exposure Time and Weapon Presence/Timing on Eyewitness Confidence and Accuracy. *Applied Cognitive Psychology*, *30(6)*, 898-910. https://doi.org/10.1002/acp.3275

*Charman, S. D., & Cahill, B. S. (2012). Witnesses' memories for lineup fillers postdicts their identification accuracy. *Journal of Applied Research in Memory and Cognition*, *1(1)*, 11-17. https://doi.org/10.1016/j.jarmac.2011.08.001

*Charman, S. D., & Quiroz, V. (2016). Blind sequential lineup administration reduces both false identifications and confidence in those false identifications. *Law and Human Behavior*, *40(5)*, 477-487. https://doi.org/10.1037/lhb0000197

Charman, S, D., & Wells, G. L. (2007). Applied lineup theory. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people*

(pp. 219–254). Mahwah, NJ: Lawrence Erlbaum and Associates.

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior*, *29(2)*, 151-172. https://doi.org/10.1007/s10979-005-2418-7

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in Eyewitness Identification. *Law and Human Behavior*, *32(3)*, 187-218. https://doi.org/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(3)*, 199-216. https://doi.org/10.1023/A:1010753809988

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair Lineups Make Witnesses More Likely to Confuse Innocent and Guilty Suspects. *Psychological Science*, *27(9)*, 1227-1239. https://doi.org/10.1177/0956797616655789

Corey, D., Malpass, R. S., & McQuiston, D. E. (1999). Parallelism in eyewitness and mock witness identifications. *Applied Cognitive Psychology*, *13(S1)*, S41-S58. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-A

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(4)*, 629-637. https://doi.org/10.1037/0021-9010.72.4.629

*Davis, J. P., Gibson, S., & Solomon, C. (2014). The Positive Influence of Creating a Holistic Facial Composite on Video Line-up Identification. *Applied Cognitive Psychology*, *28(5)*, 634-639. https://doi.org/10.1002/acp.3045

*Davis, J. P., Maigut, A. C., Jolliffe, D., Gibson, S. J., & Solomon, C. J. (2015). Holistic Facial Composite Creation and Subsequent Video Line-up Eyewitness Identification Paradigm.

Journal of Visualized Experiments, *106*, e53298. https://doi.org/10.3791/53298

*Davis, J. P., Thorniley, S., Gibson, S. J., & Solomon, C. J. (2016). Holistic facial composite construction and subsequent lineup identification accuracy: comparing adults and children. *Journal of Psychology: Interdisciplinary and Applied, 150(1)*, 102-118. https://doi.org/10.1080/00223980.2015.1009867

DeVellis, R. (2012). *Scale Development Theory and Applications*. New York, NY: Sage Publications.

Doob, A. N., & Kirshenbaum, H. M. (1973). The effects on arousal of frustration and aggressive films. *Journal of Experimental Social Psychology, 9(1)*, 57-64. https://doi.org/10.1016/0022-1031(73)90062-0

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research & Perspectives, 38(1)*, 105-124.

*Dysart, J. E., Lawson, V. Z., & Rainey, A. (2012). Blind lineup administration as a prophylactic against the postidentification feedback effect. *Law and Human Behavior*, *36(4)*, 312-319. https://doi.org/10.1037/h0093921

Engellant, K. A., Holland, D. D., & Piper, R. T. (2016). Assessing Convergent and Discriminant Validity of the Motivation Construct for the Technology Integration Education (TIE) Model. *Journal of Higher Education Theory and Practice, 16*. 37-50. https://doi.org/10.33423/jhetp.v16i1.1935

*Fitzgerald, R. J., Oriet, C., & Price, H. L. (2016). Change blindness and eyewitness identification: Effects on accuracy and confidence. *Legal and Criminological Psychology*, *21(1)*, 189-201. https://doi.org/10.1111/lcrp.12044

Fitzgerald, R. J., Oriet, C., & Price, H. L., (2015). Suspect filler similarity in eyewitness lineups:

A literature review and a novel methodology. *Law and Human Behavior*, *39(1)*, 62-74. https://doi.org/10.1037/lhb0000095

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(2)*, 151–164. https://doi.org/10.1037/a0030618

Gepshtein, S., Wang, Y., He, F., Diep, D., & Albright, T. D. (2020). A perceptual scaling approach to eyewitness identification. *Nature communications, 11(1)*, 1-10. https://www.nature.com/articles/s41467-020-17194-5

Gonzalez, R., Davis, J., & Ellsworth, P. C. (1995). Who should stand next to the suspect? Problems in the assessment of lineup fairness. *Journal of Applied Psychology, 80(4)*, 525-531. https://doi.org/10.1037/0021-9010.80.4.525

*Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior*, *33(1)*, 70-82. https://doi.org/10.1007/s10979-008-9136-x

*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(2)*, 140-152. https://doi.org/10.1037/a0015082

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research, 57*, 98-107. https://doi.org/10.1016/S0148-2963(01)00295-8.

*Haw, R. M., Dickinson, J. J., & Meissner, C. A. (2007). The phenomenology of carryover effects between show-up and line-up identification. *Memory*, *15*, 117-127. https://doi.org/10.1080/09658210601171672

*Haw, R. M., & Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *The Journal of Applied Psychology*, *89(6)*, 1106-1112. https://doi.org/10.1037/0021-9010.89.6.1106

Hirschberg, N., Jones, L. E., & Haggerty, M. (1978). What's in a face: Individual differences in face perception. *Journal of Research in Personality*, *12(4)*, 488-499. https://doi.org/10.1016/0092-6566(78)90074-0

*Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology Applied*, *18(4)*, 346-360. https://doi.org/10.1037/a0029779

*Hosch, H. M., & Bothwell, R. K. (1990). Arousal, description and identification accuracy of victims and bystanders. *Journal of Social Behavior & Personality, 5(5)*, 481-488

*Humphries, J. E., Holliday, R. E., & Flowe, H. D. (2012). Faces in Motion: Age-Related Changes in Eyewitness Identification Performance in Simultaneous, Sequential, and Elimination Video Lineups. *Applied Cognitive Psychology*, *26(1)*, 149-158. https://doi.org/10.1002/acp.1808

*Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. *Psychology, Crime & Law*, *21(9)*, 871-889. https://doi.org/10.1080/1068316X.2015.1054387

*Key, K. N., Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Cash, D. K., & Lane, S. (2017). Line-up Fairness Affects Postdictor Validity and 'Don't Know' Responses. *Applied Cognitive Psychology*, *31(1)*, 59-68. https://doi.org/10.1002/acp.3302

*Köhnken, G., & Maass, A. (1988). Eyewitness Testimony: False Alarms on Biased Instructions?

Journal of Applied Psychology, *73(3)*, 363-370. https://doi.org/10.1037/0021-9010.73.3.363

*Krafka, C., & Penrod, S. (1985). Reinstatement of Context in a Field Experiment on Eyewitness Identification. *Journal of Personality and Social Psychology*, *49(1)*, 58-69. https://doi.org/10.1037/0022-3514.49.1.58

*Lawson, V. Z., & Dysart, J. E. (2014). The showup identification procedure: An exploration of systematic biases. *Legal and Criminological Psychology*, *19(1)*, 54-68. https://doi.org/10.1111/j.2044-8333.2012.02057.x

Lee, J. (2019). Facial identification: A meta-analysis of 50 years of research (Doctoral dissertation). Retrieved from CUNY Academic Works (https://academicworks.cuny.edu/gc_etds/3422).

Lee, J., & Penrod, S. D. (2019). New signal-detection-theory-based framework for eyewitness performance in lineups. *Law and Human Behavior. 43(5)*, 436-454. https://doi.org/10.1037/lhb0000343.

Litwin, M. S. (1995). *How to measure survey reliability and validity*. CA: SAGE Publications, Inc. doi: 10.4135/9781483348957.

*Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about “weapon focus.” *Law and Human Behavior*, *11(1)*, 55-62. https://doi.org/10.1007/BF01044839

Lindsay, R. C. L., Mansour, J. K., Beaudry, J. L., Leach, A. -M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal & Criminological Psychology, 14(1)*, 13-24. https://doi.org/10.1348/135532508X382708

Lindsay, R. C. L., Smith, S. M., & Pryke, S. (1999). Measures of lineup fairness: do they postdict identification accuracy? *Applied Cognitive Psychology, 13*, S93-S107.

https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S93::AID-ACP633>3.0.CO;2-X

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(4)*, 303-313. https://doi.org/10.1007/BF01040622

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior*, *5(4)*, 299-309. https://doi.org/10.1007/BF01044945

Malpass, R. S., & Lindsay, R. C. L. (1999). Measuring lineup fairness. *Applied Cognitive Psychology*, *13(S1)*, 1-7. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-9

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 (Vol. 2): Memory for People*. Mahwah, NY: Lawrence Erlbaum & Associates.

Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I., & Lindsay, R. C. L. (2017). Evaluating lineup fairness: Variations across methods and measures. *Law and Human Behavior*, *41(1)*, 103-115. https://doi.org/10.1037/lhb0000203

*Mansour, J. K., Lindsay, R. C. L., Brewer, N., & Munhall, K. G. (2009). Characterizing visual behaviour in a lineup task. *Applied Cognitive Psychology*, *23(7)*, 1012-1026. https://doi.org/10.1002/acp.1570

Malpass, R. S. & Devine, P. G. (1983). Measuring the fairness of eyewitness identification lineups. In S. Lloyd-Bostock, & B. Clifford (Eds.), *Evaluating Witness Evidence.* (pp. 81-102). London- Wiley & Sons

McQuiston, D. E., & Malpass, R. S. (2002). Validity of the Mockwitness Paradigm: Testing the

Assumptions. *Law and Human Behavior*, *26(4)*, 439-453. https://doi.org/10.1023/A:1016383305868

*Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition*, *29(1)*, 176-186. https://doi.org/10.3758/BF03195751

Nosworthy, G. J., & Lindsay, R. C. L. (1990). Does nominal Lineup Size matter*? Journal of Applied Psychology, 75(3)*, 358-361. http://dx.doi: 10.1037/0021-9010.75.3.358

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.).* New York, NY: McGraw-Hill.

Penrod, S. D. (2003). Eyewitness identification evidence: How well are witnesses and police performing? *Criminal Justice Magazine, Spring*, 36-47.

Penrod, S. D. (2006, March). *Eyewitness guessing and choosing*. Paper presented at the meeting of the American Psychology-Law Society, St. Petersburg, FL.

*Pigott, M. A., Brigham, J. C., & Bothwell, R. K. (1990). Field study on the relationship between quality of eyewitnesses' descriptions and identification accuracy. *Journal of Police Science and Administration, 17(2)*, 84-88.

*Platz, S. J., & Hosch, H. M. (1988). Cross-Racial/Ethnic Eyewitness Identification: A Field Study1. *Journal of Applied Social Psychology*, *18(11)*, 972-984. https://doi.org/10.1111/j.1559-1816.1988.tb01187.x

Post, M. W. (2016). What to Do With "Moderate" Reliability and Validity Coefficients? *Archives of Physical Medicine and Rehabilitation, 97(7)*, 1051-1052. https://doi.org/10.1016/j.apmr.2016.04.001

Pozzulo, J. D., Dempsey, J. L., & Wells, K. (2010). Does Lineup Size Matter with Child

Witnesses. *Journal of Police and Criminal Psychology*, *25(1)*, 22-26. https://doi.org/10.1007/s11896-009-9055-x

*Pozzulo, J. D., Reed, J., Pettalia, J., & Dempsey, J. (2016). Simultaneous, Sequential, Elimination, and Wildcard: A Comparison of Lineup Procedures. *Journal of Police and Criminal Psychology*, *31(1)*, 71-80. https://doi.org/10.1007/s11896-015-9168-3

*Quinlivan, D. S., Neuschatz, J. S., Cutler, B. L., Wells, G. L., McClung, J., & Harker, D. L. (2012). Do pre-admonition suggestions moderate the effect of unbiased lineup instructions? *Legal and Criminological Psychology*, *17(1)*, 165-176. https://doi.org/10.1348/135532510X533554

*Rhead, L. M., Rodriguez, D. N., Korobeynikov, V., Yip, J. H., & Kovera, M. B. (2015). The Effects of Lineup Administrator Influence and Mortality Salience on Witness Identification Accuracy. *Journal of Forensic Psychology Practice*, *15(3)*, 248-274. https://doi.org/10.1080/15228932.2015.1041362

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shapiro, P. N. & Penrod, S. D. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin 100(2),* 139-156. https://doi.org/10.1037/0033-2909.100.2.139

*Smalarz, L., & Wells, G. L. (2014). Confirming feedback following a mistaken identification impairs memory for the culprit. *Law and Human Behavior*, *38(3)*, 283-292. https://doi.org/10.1037/lhb0000078

*Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential Lineup Laps and Eyewitness Accuracy. *Law and Human Behavior*, *35(4)*, 262-274. https://doi.org/10.1007/s10979-010-9236-2

*Steblay, N. K., & Phillips, J. D. (2011). The not-sure response option in sequential lineup practice. *Applied Cognitive Psychology*, *25(5)*, 768-774. https://doi.org/10.1002/acp.1755

Stovall v. Denno, 388 U. S. (1967)

*Sučić, I., Tokić, D., & Ivešić, M. (2015). Field study of response accuracy and decision confidence with regard to lineup composition and lineup presentation. *Psychology, Crime & Law*, *21(8),* 798-819. https://doi.org/10.1080/1068316X.2015.1054383

Tredoux, C. G. (1998). Statistical Inference on Measures of Lineup Fairness. *Law and Human Behavior*, *22(2)*, 217-237. https://doi.org/10.1023/A:1025746220886

Tredoux, C. G. (1999). Statistical considerations when determining measures of Lineup Size and lineup bias. *Applied Cognitive Psychology*, *13(S1)*, 9-26. https://doi.org/10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-1

Tredoux, C. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. *Journal of Experimental Psychology: Applied, 8(3)*, 180–193. https://doi.org/10.1037/1076-898X.8.3.180

Tredoux, C. (2019, May). *Identification parades: Measuring their fairness, creating synthetic versions and motivations for change of format in South Africa*. Invited presentation at the National VIPER© Bureau Academic Research Conference “Eyewitness Identification”, National VIPER© Bureau, Wakefield, UK.

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43(2),* 161-204. https://doi.org/10.1080/14640749108400966

*Vredeveldt, A., Tredoux, C. G., Kempen, K., & Nortje, A. (2015). Eye Remember What Happened: Eye-Closure Improves Recall of Events but not Face Recognition. *Applied*

Cognitive Psychology, *29(2)*, 169-180. https://doi.org/10.1002/acp.3092

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, *3(4)*, 285-293. https://doi.org/10.1007/BF01039807

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99(3)*, 320-329. http://dx.doi: 10.1037/0033-2909.99.3.320

Tables

Table 1

Variable List of Influences on Eyewitness Performance (Memory Strength and Lineup Fairness)

Variable family	Variable		
	Memory Strength	Lineup Fairness	
		TP lineups	TA lineups
Multi-d model	$d'(R_aR_p)$ $= z(R_a) - z(R_p)$	$d'(GF_p) - d'(R_aR_p)$ $= \{z(G) - z(F_p)\} -$ $\{z(R_a) - z(R_p)\}$	$d'(IF_a)$ $= z(I) - z(F_a)$
Strategy-based breakdown	Reliable witnesses $= R_a - R_p$	TP bias $= (G - (R_a - R_p))$ $/(\text{TP ID}/N)$	TA bias $= I/(\text{TA ID}/N)$
Study characteristics	Exposure time	Filler similarity	Filler similarity

Note: G = guilty suspect identification rate, F_p = filler identification rate in a TP lineup, R_p = rejection rate in a TP lineup, I = innocent suspect identification rate, F_a = filler identification rate in a TA lineup, R_a = rejection rate in a TA lineup, z(p) = the inverse cumulative distribution function of a normal distribution, TP ID = the total identification rate from a TP lineup, TA ID = the total identification rate from a TA lineup, and N = nominal lineup size.

Table 2.

Example Lineup Identification Rates (a 6-person lineup)

Eyewitness Response	TP lineup	TA lineup
Suspect ID	.50	.10
Filler ID	.20	.35
Rejection	.30	.55

Table 3.

Regression Analyses of Eyewitness Performance on Target Presence and Lineup Size

DV	Predictor	*B*	Std. Error	Beta	*t*	P value	R Square	*F*
Suspect ID	(Constant)	.254	.027		9.288	<.001		
	Target Presence	.198	.033	.519	5.940	<.001		
	Lineup Size	-.055	.028	-.289	-1.926	.057		
	TP × LS	.035	.035	.151	1.007	.316	.305	$F(3,91) = 13.33, p < .001$
	Simple slope of Lineup Size							
	TP lineup	-.020	.020		-.962	.339		
	TA lineup	-.055	.028		-1.926	.057		
Filler ID	(Constant)	.306	.027		11.538	<.001		
	Target Presence	-.036	.033	-.110	-1.097	.276		
	Lineup Size	.100	.028	.620	3.635	<.001		
	TP × LS	-.074	.034	-.371	-2.173	.033	.163	$F(3,84) = 5.46, p = .002$
	Simple slope of Lineup Size							
	TP lineup	.026	.020		1.325	.189		
	TA lineup	.100	.028		3.635	<.001		
Rejection	(Constant)	.438	.021		21.013	<.001		
	Target Presence	-.155	.026	-.537	-5.966	<.001		
	Lineup Size	-.045	.022	-.322	-2.095	.039		
	TP × LS	.039	.027	.222	1.444	.153	.321	$F(3,84) = 13.21, p < .001$
	Simple slope of Lineup Size							
	TP lineup	-.007	.016		-.436	.664		
	TA lineup	-.045	.022		-2.095	.039		
Discriminability	(Constant)	-.171	.158		-1.088	.280		
	Target Presence	.723	.196	.349	3.695	<.001		
	Lineup Size	-.562	.164	-.553	-3.427	.001		
	TP × LS	.382	.202	.305	1.890	.062	.252	$F(3,84) = 9.45, p < .001$
	Simple slope of Lineup Size							
	TP lineup	-.180	.118		-1.522	.132		
	TA lineup	-.562	.164		-3.427	.001		

Note. TP × LS represents the interaction term of Target Presence × Lineup Size. Lineup Size was centered in the regression models.

Table 4.

MTMM Correlation Matrix for Mock-witness Measures and the Lineup Fairness and Memory Quality Indices of Other Methods

			Mock-witness		Multi-d′		Strategy-based Breakdown		Study Characteristics	
			Fairness	Memory	Fairness	Memory	Fairness	Memory		Fairness
method	trait	index	Lineup Size	Functional Size	$d'(R_aR_p)$	$d'(GF_p) - d'(R_aR_p)$, & $d'(IF_a)$	Reliable Eyewitnesses	TP & TA bias	Exposure Time	Filler Similarity
Mock-witness	Fairness	Lineup Size	1.00							
		Functional Size	.22 (24)	1.00						
Multi-d′	Memory	$d'(R_aR_p)$	.06 (57)	.07 (8)	1.00					
	Fairness	$d'(GF_p) - d'(R_aR_p)$, & $d'(IF_a)$	-.45*** (66)	-.75* (10)	-.02 (57) [-0.10, 0.09]	1.00				
Strategy-based Breakdown	Memory	Reliable Eyewitnesses	.03 (57)	.03 (8)	>.99*** (57)	.003 (57)	1.00			
	Fairness	TP & TA Bias	-.26* (66)	-.65* (10)	.05 (57)	.93*** (66)	.06 (57)	1.00		
Study Characteristics	Memory	Exposure Time	-.18† (95)	-.20 (26)	.23† (57)	-.05 (66)	.19 (57)	.04 (66)	1.00	
	Fairness	Filler Similarity	.30*** (95)	n/a	-.06 (57)	-.33** (66)	-.01 (57)	-.34** (66)	<.001 (97)	1.00

Note. represents heterotrait-heteromethod cells (i.e., divergent validity cells); represents monotrait-heteromethod cells (i.e., convergent validity cells). *** $p < .001$, ** $p < .01$, * $p < .05$, and † $p < .10$. n/a indicates that the correlation analysis was not conducted because of the small number of cases. N is in parentheses.

Figures

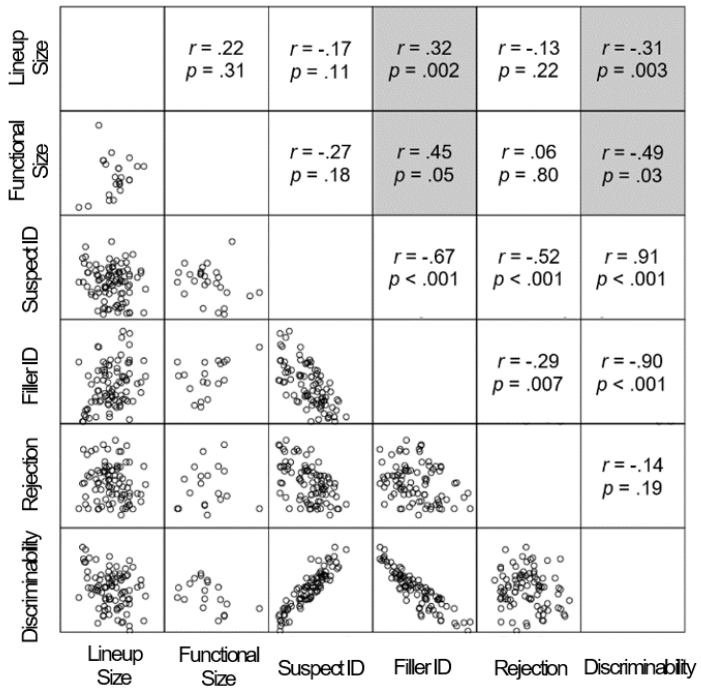

Figure 1. Correlations between mock-witness measures (Lineup Size and Functional Size) and eyewitness performance (suspect and filler IDs, rejections, and the discriminability of a suspect from fillers). Panels showing significant correlations of interest were colored as gray.

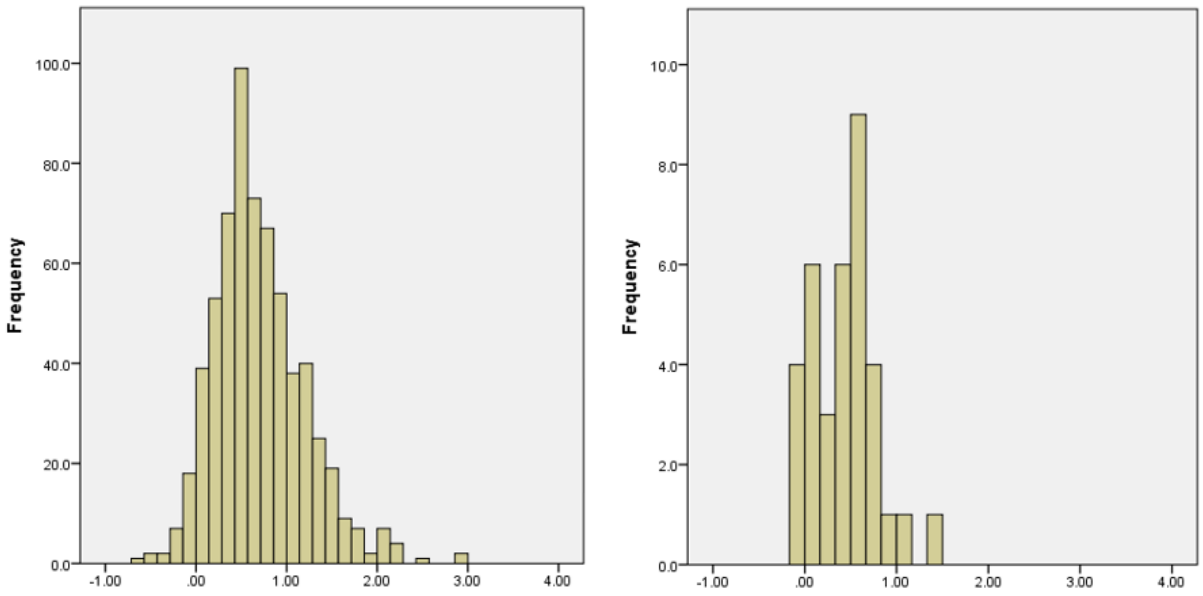

Figure 2. Histogram of d′(R_aR_p) in Lee's (2019) meta-analysis database (left) and the current database (right).

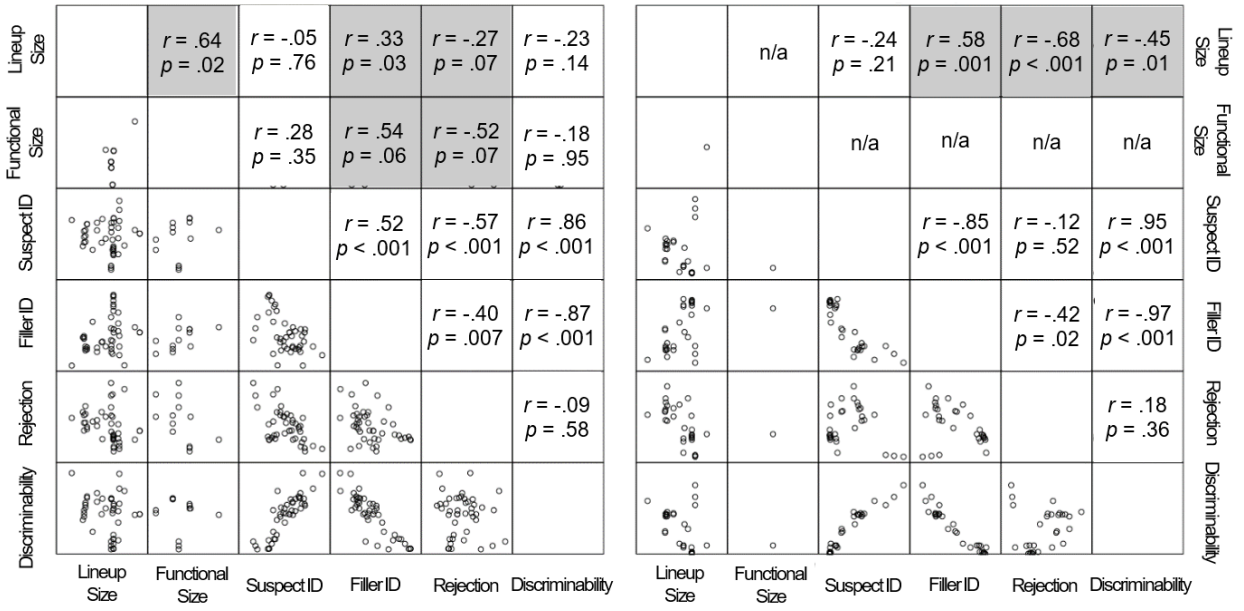

Figure 3a. Correlations and scatter plots of mock-witness measures and eyewitness performance in the weak memory condition. The left panel depicts TP lineups while the right panel depicts TA lineups. n/a indicates that there were too few observations to permit the correlation analysis.

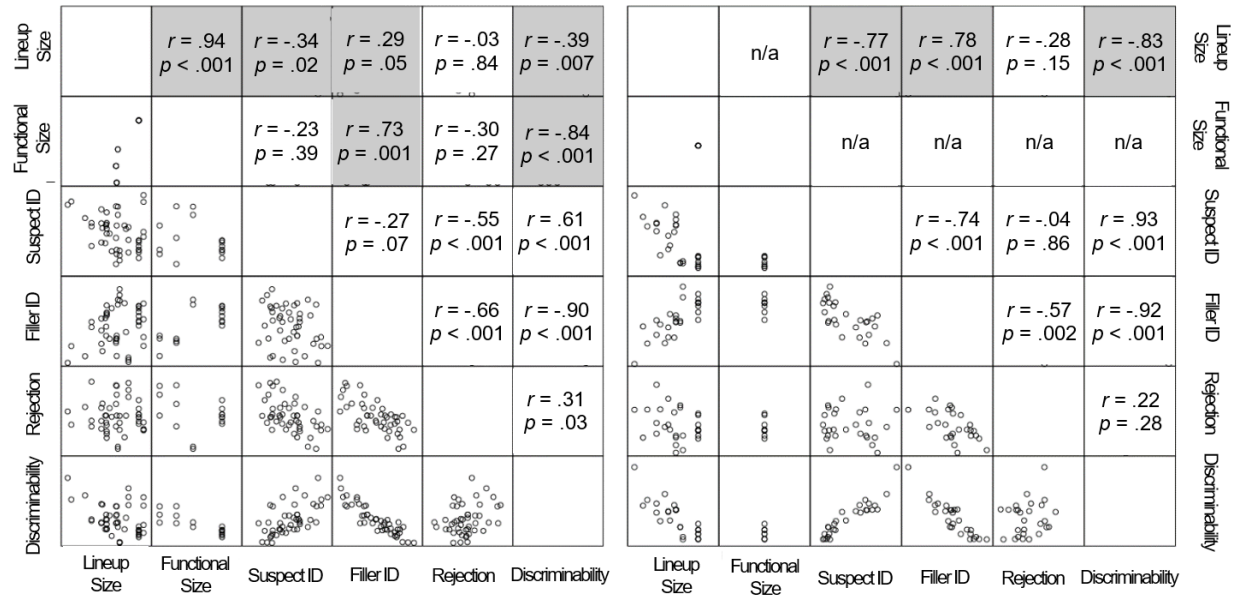

Figure 3b. Correlations and scatter plots of mock-witness measures and eyewitness performance in the moderate memory condition. The left panel depicts TP lineups while the right panel depicts TA lineups. n/a indicates that there were too few observations to permit the correlation analysis.

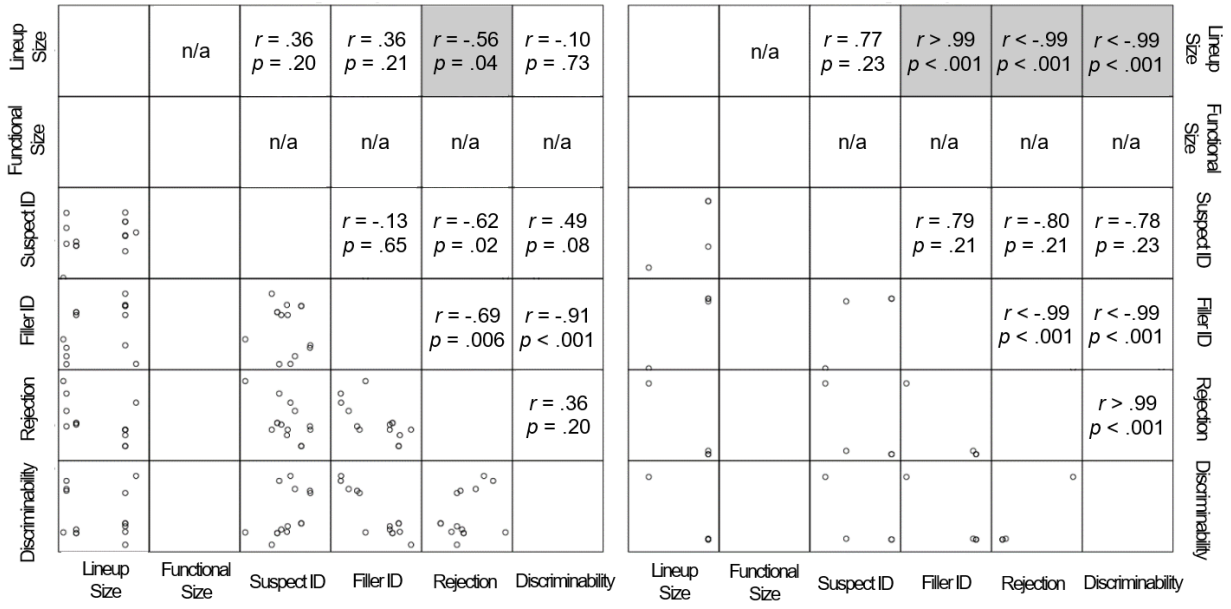

Figure 3c. Correlations and scatter plots of mock-witness measures and eyewitness performance in the strong memory condition. The left panel depicts TP lineups while the right panel depicts TA lineups. n/a indicates that there were too few observations to permit the correlation analysis.

Appendix

Overall estimates of mock-witness measures and eyewitness performance for each of the 43 studies included in the current database

Study	Target Presence	Lineup Type	Tredoux' s E	Effective Size	Functional Size	Suspect ID rate	Filler ID rate	Rejection rate
Andersen, Carlson, Carlson, & Gronlund (2014)	TP		3.69			0.56	0.14	0.31
	TA		3.50			0.37	0.21	0.43
Beresford & Blades (2006)	TP		6.67	6.91	6.25	0.45	0.30	0.26
	TA		6.00	6.82	6.25	0.07	0.47	0.46
Brigham, Maass, Snyder, & Spaulding (1982)	TP	average		3.13	6.52	0.34	0.45	0.21
	TP	Lineup1		5.01	7.33	0.29	0.48	0.24
	TP	Lineup2		4.50	20.20	0.17	0.69	0.14
	TP	Lineup3		3.40	1.88	0.27	0.36	0.36
	TP	Lineup4		3.39	2.60	0.48	0.26	0.26
	TP	Lineup5		3.38	-	0.39	0.43	0.17
	TP	Lineup6		2.58	11.50	0.29	0.57	0.14
	TP	Lineup7		1.75	1.14	0.50	0.36	0.14
	TP	Lineup8		1.00	1.00	0.40	0.46	0.14
Brigham, Verst, & Bothwell (1986)	TP			3.59	8.00	0.83		
Carlson & Carlson (2014)	TP		4.98			0.28	0.51	0.21
	TA		4.88			0.07	0.67	0.26
Carlson, Dias, Weatherford, & Carlson (2017)	TP		5.36			0.40	0.35	0.25
	TA		4.25			0.13	0.64	0.23
Carlson, Gronlund, & Clark (2008) Exp.1	TP		3.78			0.65	0.05	0.30
Carlson, Gronlund, & Clark (2008) Exp.2	TP	average	2.77			0.43	0.16	0.41
	TA	average	3.02			0.34	0.21	0.45
	TP	Lineup1	2.56			0.33	0.24	0.43
	TA	Lineup1	2.92			0.33	0.20	0.46
	TP	Lineup2	1.69			0.58	0.04	0.38
	TA	Lineup2	1.54			0.50	0.11	0.39
	TP	Lineup3	4.05			0.36	0.21	0.43
	TA	Lineup3	4.59			0.18	0.33	0.49
Carlson, Young, Weatherford, Carlson, Bednarz, & Jones (2016)	TP		4.07			0.56	0.19	0.25
	TA		4.65			0.06	0.32	0.61
Charman & Cahill (2012)	TP		4.00			0.46	0.24	0.30
Charman, & Quiroz (2016)	TP		5.45			0.68	0.18	0.14
	TA		5.13			0.69	0.17	0.14
Davis, Gibson, & Solomon (2014)	TP		7.05			0.50	0.31	0.20
	TA		4.85			0.07	0.59	0.34
Davis, Maigut, Jolliffe, Gibson, & Solomon (2015)	TP		7.05			0.52	0.23	0.25
Davis, Thorniley, Jolliffe, Gibson, & Solomon (2016)	TP		7.53			0.32	0.35	0.32
Dysart, Lawson, & Rainey (2012)	TP		4.26			0.31		
Fitzgerald, Oriet, & Price (2016)	TP		3.34			0.29	0.36	0.35
	TA		3.94			0.08	0.47	0.45
Greathouse & Kovera (2009)	TP		5.00		4.82	0.60	0.30	0.10
Gronlund, Carlson, Dailey, & Goodsell (2009)	TP	average	2.65			0.42	0.25	0.34
	TA	average	2.91			0.35	0.23	0.42
	TP	Lineup1	4.51			0.63	0.12	0.25
	TP	Lineup2	3.75			0.11	0.43	0.46
	TP	Lineup3	2.57			0.09	0.58	0.33
	TP	Lineup4	2.33			0.52	0.20	0.29
	TP	Lineup5	1.47			0.35	0.10	0.55
	TP	Lineup6	1.29			0.77	0.05	0.18
	TA	Lineup7	4.35			0.51	0.18	0.31
	TA	Lineup8	3.88			0.15	0.41	0.44
	TA	Lineup9	3.15			0.46	0.15	0.40
	TA	Lineup10	2.91			0.13	0.38	0.49

	TA	Lineup11	1.85			0.26	0.16	0.57
	TA	Lineup12	1.29			0.60	0.07	0.33
Haw, Dickinson, & Meissner (2007)	TP		4.99			0.42		
Horry, Palmer, & Brewer (2012)	TP		3.69			0.47	0.18	0.36
	TA		3.75			0.13	0.21	0.65
Hosch, & Bothwell (1990)	TP			4.15	6.00	0.50		
Humphries, Holliday, & Flowe (2012)	TP		5.53			0.61	0.24	0.16
Key, Cash, Neuschatz, Price, Wetmore, & Gronlund (2015)	TP	average	2.90			0.54	0.17	0.29
	TA	average	2.87			0.19	0.40	0.41
	TP	Lineup1	4.51			0.42	0.29	0.29
	TA	Lineup1	3.88			0.15	0.39	0.46
	TP	Lineup2	1.29			0.67	0.10	0.23
	TA	Lineup2	1.85			0.29	0.29	0.41
Key, Wetmore, Neuschatz, Gronlund, Cash, & Lane (2017)	TP	average	2.39			0.59	0.05	0.36
	TA	average	2.18			0.25	0.17	0.59
	TP	Lineup1	1.00			0.65	0.02	0.33
	TA	Lineup1	1.21			0.40	0.02	0.59
	TP	Lineup2	3.77			0.54	0.07	0.39
	TA	Lineup2	3.15			0.10	0.31	0.59
Kohnken & Maass (1988) Exp.1	TA			4.67	7.00	0.05	0.33	0.62
Kohnken & Maass (1988) Exp.2	TA			4.67	7.00	0.12	0.46	0.42
Krafka & Penrod (1985)	TP			4.29	3.18	0.41	0.14	0.46
Lawson & Dysart (2014)	TP		4.34			0.51	0.14	0.36
	TA		4.20			0.22	0.21	0.57
Loftus, Loftus, & Messo (1987)	TP				10.29	0.25		
Mansour, Lindsay, Brewer, & Munhall (2009)	TP			3.87	4.87	0.40	0.18	0.39
Meissner, Brigham, & Kelley (2001) Exp.1	TP			5.00	4.50	0.46	0.37	0.17
Meissner, Brigham, & Kelley (2001) Exp.2	TP			5.00	4.50	0.37	0.46	0.17
Pigott, Brigham, & Bothwell (1990)	TP	average		3.95	3.96	0.48		
	TP	Lineup1		4.15	6.00	0.39		
	TP	Lineup2		3.75	1.92	0.60		
Platz &Hosch (1988)	TP	average		2.82		0.44	0.40	0.16
	TP	Lineup1		3.50		0.38	0.48	0.14
	TP	Lineup2		2.52		0.47	0.40	0.14
	TP	Lineup3		2.44		0.48	0.33	0.20
Pozzulo, Reed, Pettalia, & Dempsey (2016)	TP		5.65			0.53	0.17	0.30
Quinlivan, Neuschatz, Cutler, Wells, McClung, & Harker (2012)	TA		4.75			0.35	0.45	0.19
Rhead, Rodriguez, Korobeynikov, Yip, & Kovera (2015)	TP		4.11			0.56	0.26	0.18
	TA		4.33			0.28	0.35	0.37
Smalarz & Wells (2014)	TP				4.20	0.56		
Steblay, Dietrich, Ryan, Raczynski, & James (2011) Exp.1	TP		4.79		4.29	0.12	0.31	0.57
Steblay, Dietrich, Ryan, Raczynski, & James (2011) Exp.2	TP		4.81		4.00	0.51	0.16	0.33
Steblay & Phillips (2011)	TP		4.86		4.00	0.49	0.13	0.39
Sučić, Tokić, & Ivešić (2015)	TP		5.14			0.40	0.35	0.25
Vredeveldt, Tredoux, Kempen,& Nortje (2015)	TP		5.85			0.48	0.03	0.49

Note. When the authors used multiple lineups with different perpetrators in a study, we included estimates associated with each of the lineups as well as the average estimates. TP: target-present lineups and TA: target-absent lineups. The data is available from the corresponding author upon reasonable request.