Estimation of Eyewitness Error Rates in Fair and Biased Lineups

Ryan J. Fitzgerald[1], Colin G. Tredoux[2], & Stefana Juncu[3]

[1]Department of Psychology, Simon Fraser University
[2]Department of Psychology, University of Cape Town
[3]Department of Psychology, University of Portsmouth

Stimuli and data are available at https://osf.io/b7gsu/?view_only=6bfaed7f72ac4bdca25a19872c15f93f.

Accepted for publication in Law and Human Behavior.

Experiment 1 was supported by a grant from the British Academy to Ryan J. Fitzgerald (SG161889). Experiment 2 was supported by a Future Research Leaders grant from the Economic and Social Research Council to Ryan J. Fitzgerald (ES/N016602/1). Experiment 3 was supported by an Insight Grant from the Social Sciences and Humanities Council of Canada to Ryan J. Fitzgerald (435-2021-1199).

Correspondence concerning this article should be addressed to Ryan J. Fitzgerald, Department of Psychology, Simon Fraser University, 8888 University Drive, Burnaby, B.C., Canada V5A 1S6. Email: r_fitzgerald@sfu.ca

Abstract

Objective: The risk of mistaken identification for innocent suspects in lineups can be estimated by correcting the overall error rate by the number of people in the lineup. We compared this nominal size correction to a new effective size correction, which adjusts the error rate for the number of plausible lineup members.

Hypotheses: (1) Increasing lineup bias will increase misidentifications of a designated innocent suspect; (2) With the effective size correction, increasing lineup bias will also increase the estimate of innocent suspect misidentifications; (3) With the nominal size correction, lineup bias will have no effect on the estimate of innocent suspect misidentifications.

Method: In a re-analysis of previous literature, we obtained 10 datasets from the Open Science Framework. In three new experiments (N_1 = 686, N_2 = 405, N_3 = 1531), participants observed a staged crime and completed a fair or biased lineup.

Results: In the re-analysis of previous literature, less than 4 out of 6 lineup members were identified frequently enough to be classified as plausible (M = 3.78, 95% CI [2.20, 5.36]). In the new experiments, increasing lineup bias increased mistaken identifications of a designated innocent suspect, OR = 5.50, 95% CI [2.77, 10.95], and also increased the effective-size-corrected estimate of innocent suspect misidentifications, OR = 3.04, 95% CI [2.13, 4.33]. With the nominal size correction, lineup bias had no effect on the estimate of innocent suspect misidentifications, OR = 0.84, 95% CI [0.60, 1.18].

Conclusion: Most lineups include a combination of plausible and implausible lineup members. Contrary to the nominal size correction, which ignores implausible lineup members, the effective size correction is sensitive to implausible lineup members and accounts for lineup bias when estimating the risk to innocent suspects.

Keywords: Eyewitness identification, error rate, lineup bias, lineup fairness, confidence

Public Significance Statement

An increasingly common method of estimating the risk to innocent suspects in fair lineups is to divide the overall error rate by the number of people in the lineup. An alternative method is to estimate the risk to innocent suspects in lineups, given the fairness of the lineups that were tested. Neither method applies to all criminal investigations, and we recommend reporting both methods when estimating the risk to innocent suspects in police lineups.

Estimation of Eyewitness Error Rates in Fair and Biased Lineups

A new perspective is emerging on the reliability of eyewitness identification (ID) from police lineups. With the rise of post-conviction DNA testing, mistaken eyewitness ID has become known as a leading contributing factor in wrongful convictions (Innocence Project, 2023; Wells et al., 1998). Eyewitness expressions of confidence have also been described as an imperfect indicator of ID accuracy (Penrod & Cutler, 1995; Sporer et al., 1995). Conversely, in recent literature Wixted and colleagues have proposed that eyewitness IDs are a reliable form of evidence so long as the eyewitness is highly confident, the lineup procedure adheres to best practice, and there has been no pre-lineup contamination of the eyewitness' memory (Wixted, 2018; Wixted et al., 2015, 2018). In support of this new characterization of eyewitness reliability, Wixted and Wells (2017) analyzed experimental studies with "pristine" lineup procedures (Table 1) and estimated a 3% error rate for eyewitnesses who identified the suspect with high confidence. When estimating the error rate, however, Wixted and Wells made an assumption: They assumed the lineups were fair.

In modern lineup studies, error rate estimates typically include only a fraction of the ID errors that occurred in the experiment. This is because when the police assemble a lineup, best practice is to use the single-suspect model (Wells et al., 1998, 2020). In this model only one lineup member is suspected to be the culprit, and the rest of the lineup members are known-innocent fillers. If the police are investigating an innocent suspect and an eyewitness mistakenly identifies an innocent suspect, it could result in a wrongful conviction. When a filler is misidentified, however, the police know the filler is innocent and that the eyewitness made an error. Therefore, although filler IDs can impeach the credibility of the eyewitness (Smalarz et al., 2019), they do not pose a risk of wrongful conviction (Wells & Turtle, 1986). Accordingly,

researchers typically classify some portion of the eyewitness errors as fillers IDs, and those errors are excluded from estimates of the suspect ID error rate.

There is more than one way to estimate the rate of innocent suspect IDs in experimental data. Wixted and Wells (2017) used a *nominal lineup size correction*, which divides the overall mistaken ID rate in a culprit-absent lineup by the number of lineup members. The nominal size correction ignores differences in the number of IDs attracted by each lineup member and, hence, assumes the lineup is fair. Another method of estimating the innocent suspect ID rate is the *effective size correction* (Smith et al., 2021). This new method corrects for the number of plausible or "effective" lineup members by accounting for how the eyewitness IDs distribute across the lineup. Instead of assuming the lineup is fair, the effective size correction assumes the innocent suspect is among the plausible lineup options. In this paper, we introduce a method for depicting how these assumptions affect estimates of the suspect ID error rate.

Classification of Innocent Suspect IDs: Designation vs Estimation

To calculate the suspect ID error rate, guilty suspect IDs and innocent suspect IDs are both needed. A guilty suspect ID occurs when the culprit is present in the lineup and the eyewitness correctly identifies that culprit. In experiments, culprit-present lineups are simulated by staging a crime for witnesses and then presenting a lineup that includes the culprit. Because the witnessed event was staged, culprit-present lineup choices can be objectively classified as culprit IDs or filler IDs.

Classification of IDs from culprit-absent lineups is less straightforward. In a criminal investigation, the culprit would be absent from the lineup if the police were investigating an innocent suspect. This can be simulated in experiments by constructing a lineup that does not contain the culprit from the staged crime; however, contrary to a real police lineup, in

experiments none of the culprit-absent lineup members is under any sort of investigation. Accordingly, researchers typically classify errors in culprit-absent lineups using one of two approaches: (a) *designation,* in which IDs of one culprit-absent lineup member are classified as innocent suspect IDs and the rest are classified as filler IDs, or (b) *estimation,* in which a correction is made to the overall number of culprit-absent lineup IDs to classify a fraction of them as innocent suspect IDs.

The traditional method of designation is the culprit-replacement strategy. In this method, researchers start by creating a lineup that contains the culprit and a set of culprit-matched fillers. Next, they create a culprit-absent lineup by replacing the culprit with a new lineup member to serve as the a priori designated innocent suspect. The culprit-replacement strategy has high internal validity because the same fillers are used in culprit-present and culprit-absent lineups, and the only difference across lineups is the identity of the suspect (Wixted & Wells, 2017). However, the ecological validity of this method is questionable because the fillers in the culprit-absent lineup have been effectively matched to the culprit. If the police were investigating an innocent suspect in a criminal case, the appearance of the true culprit would not be known and fillers in the culprit-absent lineup would normally be matched instead to the innocent suspect (Clark & Tunnicliff, 2001). Alternative methods of designation have been developed to improve ecological validity, such as using a single lineup and manipulating the suspect's guilt by randomly assigning participants to witness different culprits (Oriet & Fitzgerald, 2018). Post-hoc designation is also possible, such as using the most frequently chosen lineup member as the innocent suspect to simulate a worst-case scenario (Pryke et al., 2004).

One limitation of designation is that, unlike the guilty suspect, who is always the person who committed the crime at the witnessed event, in experiments there is often no justification for

designating one lineup member as the innocent suspect over any of the others. Designation is usually appropriate for studying factors that have a disproportionate effect on innocent suspect IDs, which are known as suspect bias variables (Wells & Olson, 2001). For example, innocent suspects are more likely to be identified from 3-member lineups than from 6-member lineups (Juncu & Fitzgerald, 2021), and this effect can be produced by designating an innocent suspect and varying how many fillers are added to the lineup. Wells and Olson distinguish suspect bias variables from general impairment variables, which affect eyewitness accuracy but do not have a disproportionate effect on innocent suspect IDs. When studying general impairment variables, the designation of a suspect is often arbitrary. For example, if the only experimental manipulation is the length of time that the witness views the culprit, there would be no reason to designate one lineup member as the innocent suspect over the others.

A further limitation of the designation approach is that the researcher decides how much the innocent suspect should resemble the culprit. Regardless of whether suspect bias or general impairment variables are studied, designation always requires the researcher to choose a designated innocent suspect and, depending on how closely their choice resembles the culprit, their selection could have a substantial influence on the rate of innocent suspect IDs. Quigley-McBride and Wells (2021) recommend counterbalancing the identities of the lineup members and randomly assigning an innocent suspect for each trial in the experiment. This method can be implemented by first constructing a culprit-absent lineup and then rotating which lineup member is replaced with the culprit on culprit-present trials. This is an effective strategy for equating bias across culprit-present and culprit-absent lineups, which is crucial for interpreting absolute rates of suspect ID accuracy. However, this strategy is intended for research questions that call for a

fair lineup and would not be feasible if the goal were, for example, to compare relative differences in the suspect ID rate between fair and biased lineups.

Rather than designating an innocent suspect, it has become increasingly common to estimate the number of innocent suspect IDs using the nominal size correction. By correcting the total number of culprit-absent lineup IDs by number of lineup members, the nominal lineup size correction assumes the innocent suspect attracts no more IDs than does the average filler. Thus, similar to how fair lineups can be simulated by designating the innocent suspect at random across trials (Quigley-McBride & Wells, 2021), the nominal size correction provides a statistical method of estimating the risk to innocent suspects in fair lineups.

The main limitation of the nominal size correction is its assumption that innocent suspects are no more likely than lineup fillers to be mistakenly identified. This assumption would be violated if the conditions of the eyewitness ID are not pristine (Table 1). For instance, if the lineup administrator knows which lineup member is the suspect and leaks cues to the suspect's identity to the eyewitness, the biased procedure would elevate the risk of an innocent suspect ID (Kovera & Evelo, 2017, 2021). Although the prevalence of pristine ID conditions in criminal investigations is uncertain (Smalarz & Wells, 2015), even if all the recommended practices are followed it would not negate the fact that suspects and fillers are in the lineup for different reasons. Suspects are there because the police have some reason to investigate them (Hyman, 2021). By contrast, the defining characteristic of fillers in the single-suspect model is that they are known to be innocent and not under investigation. The nominal size correction would only be applicable if being a suspect has no effect on a lineup member's risk of mistaken ID.

The choice between designation or estimation has consequences for interpreting experimental data. For instance, Lindsay and Wells (1985) reported that eyewitnesses were

significantly less likely to misidentify a designated innocent suspect in sequential lineups (17%) than in simultaneous lineups (43%). But if instead of designation Lindsay and Wells had applied the nominal size correction, they would have estimated a more modest reduction in innocent suspect IDs for sequential (6%) relative to simultaneous (10%) presentation. Thus, a 26% difference in designated innocent suspect IDs would be reduced to a 4% difference in estimated innocent suspect IDs. In the next section, we discuss the possibility that these disparate outcomes are representative of different types of eyewitness ID cases (Lee & Penrod, 2019).

How Reason of Suspicion Affects the Risk to Innocent Suspects

Lee and Penrod (2019) reviewed studies that did or did not have a designated innocent suspect. In studies with no designated innocent suspect, the nominal size correction was applied. Consistent with a previous review of the literature (Clark et al., 2008), the nominal size correction resulted in estimates of the innocent suspect ID rate that were lower than the rates observed in studies with a designated innocent suspect. Lee and Penrod noted that in many of the designation studies the researchers intentionally constructed the lineups to be biased against the innocent suspect, leading them to conclude that designation and estimation studies would simulate innocent suspects in real-world cases who are under investigation for different reasons.

Designation studies were viewed as most applicable to innocent suspects under investigation for *appearance-based reasons of suspicion* (Lee & Penrod, 2019). Consider, for example, the commonly cited wrongful conviction of Ronald Cotton. Cotton became a suspect because someone saw a composite sketch of the true culprit and notified the police that Cotton might be the man in the sketch. Postconviction DNA testing eventually implicated another man, Bobby Poole, who confessed that he was the true culprit. Cotton and Poole were similar in appearance, so similar in fact that when they were incarcerated in the same prison they were

often mistaken for one another. The similarity between Cotton and Poole was no coincidence – matching a composite sketch is an appearance-based reason of suspicion, and Cotton became a suspect because of his similarity to the sketch of Poole. Thus, appearance-based reasons of suspicion increase the risk that an innocent suspect will resemble the actual culprit (Wells & Penrod, 2011).

Given that fillers and innocent suspects are assumed to be equally plausible with the nominal size correction, Lee and Penrod (2019) proposed that this approach would be more applicable to *non-appearance-based reasons of suspicion*, which do not increase the risk that an innocent suspect would strongly resemble the true culprit (Wells & Penrod, 2011). Beyond the reason of suspicion, the equal plausibility assumption would also require that the lineup conditions are otherwise pristine. Essentially, the nominal size correction gives an estimate of the risk to innocent suspects under ideal circumstances. In the next section we argue that it is also desirable to estimate the risk to innocent suspects under less favorable conditions.

Estimating the Positive Predictive Value (PPV) of Suspect IDs

When triers of fact assess the reliability of an eyewitness ID, they need to know the suspect ID accuracy rate for whatever level of confidence was reported by the eyewitness in their case (Mickes, 2015). Throughout our article, we operationalize accuracy as the Positive Predictive Value (PPV) of suspect IDs (Mickes, 2016), which is computed by dividing the number of guilty suspect IDs by the total number of suspect IDs (i.e., guilty suspects + innocent suspects). In medical diagnostic tests, PPV is the probability that the disease is present if the diagnostic test comes back positive. For lineup responses, PPV is the probability that the suspect is guilty if the eyewitness identifies them from the lineup.

Mickes (2015) recommends grouping participants by confidence and plotting PPV for each group on a Confidence Accuracy Characteristic (CAC) curve. This approach was applied by Wixted and Wells (2017), who estimated PPV under pristine or non-pristine testing conditions in their re-analysis of the literature. Under non-pristine conditions, Wixted and Wells found that confidence was a weak predictor of accuracy. However, in the 15 studies they classified as having pristine conditions, suspect IDs made with 90% confidence or greater were estimated to have a PPV of 97% (Figure 1, Panel A).

Wixted and Wells (2017) also computed what is known as chooser calibration (Figure 1, Panel B). Calibration is a measure of the relation between confidence and accuracy (Brier, 1950; Cutler, & Penrod, 1989). In a chooser calibration curve, accuracy rates for eyewitnesses who choose from the lineup are plotted in relation to confidence ratings (Brewer et al., 2002; Brewer & Wells, 2006; Juslin et al., 1996). Contrary to CAC curves, which represent PPV after excluding filler IDs, chooser calibration curves include every mistaken ID from the culprit-absent lineup.

Scholars have conceptualized chooser calibration and CAC curves as qualitatively distinct measures that serve different purposes, a perspective succinctly summarized by Wixted and Wells (2017):

> Unlike a calibration curve, a [CAC] plot provides the information that judges and juries want to know when they are trying to assess the reliability of an eyewitness who identified a suspect from a lineup... A calibration curve is a perfectly appropriate way to represent the relevant data when the question concerns the confidence-accuracy relationship from the witness's perspective… However, the legal system is concerned with a different issue. (p. 24)

Another way to conceptualize CAC and calibration curves is to think of them as PPV at the polar extremes of the lineup fairness continuum. A CAC curve gives an estimate of PPV assuming the innocent suspect is no more likely to be identified than any other lineup member, which is representative of cases with perfectly fair lineups. Although calibration was never meant to be interpreted as a measure of PPV, it is computationally the same as estimating PPV with no correction to the overall mistaken ID rate. In this conceptualization every mistaken ID from a culprit-absent lineup is treated as an innocent suspect ID, which is representative of cases with lineups that are biased against an innocent suspect or lineups that deviate from the single-suspect model and consist entirely of innocent suspects with no known-innocent fillers.

In Panel C of Figure 1 we report the CAC and calibration curves together to depict what we refer to as the *PPV range*. Given that these measures have been previously understood to serve qualitatively distinct purposes, they were depicted by Wixted and Wells (2017) in separate graphs. By plotting the PPV range, with both curves in a single graph, we illustrate how the accuracy of suspect IDs depends on the assumptions made when estimating the number of innocent suspect IDs. If fair lineups are assumed, PPV is estimated to be at the top of the range. If biased lineups are assumed, it is at the bottom. Thus, the CAC curve represents PPV under a best-case scenario and the calibration curve represents PPV under a worst-case scenario.

In the following section, we consider a new method of estimating PPV that corrects for the number of plausible lineup members. Invariably, this alternative estimate of PPV lies in between the extreme limits of CAC and calibration curves.

Effective Size Correction

In addition to the nominal size correction, it is also possible to estimate the risk to innocent suspects by correcting for the lineup's effective size (Smith et al., 2021). Effective size

refers to the number of plausible lineup members (Malpass, 1981), which can be measured using Tredoux's (1998) method (see Effective Size Calculations in online Supplemental Materials). A lineup member is classified as plausible if they attract a certain number of IDs. Thus, data must be generated to produce a distribution of IDs across the lineup members. To measure the effective size of a lineup from a criminal investigation, the required data can be produced by recruiting non-witness participants and instructing them to identify the lineup member who best matches the eyewitness description of the culprit (Doob & Kirshenbaum, 1973). In eyewitness ID experiments, effective size can be measured directly from the distribution of mistaken IDs in the culprit-absent lineup (Quigley-McBride & Wells, 2021). This measure of effective size can then be used in place of the nominal lineup size to correct the overall rate of misidentification and estimate the risk to innocent suspects (Smith et al., 2021).

To illustrate how the effective size correction works, Table 2 depicts hypothetical data from 6-member culprit-absent lineups that are fair, partially biased, or maximally biased against one lineup member (#3). The effective size correction is computed by dividing the overall number of misidentifications by the lineup's effective size. In the fair lineup (Lineup A), which has 60 mistaken IDs distributed evenly across the six lineup members, the effective size correction classifies 10 as innocent suspect IDs. In the partially biased lineup (Lineup B), which has 60 mistaken IDs concentrated on three lineup members, the effective size correction adjusts and classifies 20 as innocent suspect IDs. In the maximally biased lineup (Lineup C), all the misidentifications are concentrated on #3 and the effective size correction accounts for this by classifying all misidentifications as innocent suspect IDs. Contrary to the effective size correction, which is sensitive to the distribution of lineup choices across the three lineups, Table

2 shows that lineup fairness has no effect on the number of innocent suspect IDs when classified with either the nominal size correction (always 10) or no correction (always 60).

All methods of estimation make assumptions about the risk to innocent suspects, and the applicability of those assumptions depends on the investigative scenario. The nominal size correction assumes an innocent suspect is no more likely than a filler to be among the plausible lineup options, which could be representative of criminal cases with non-appearance-based reasons of suspicion (Lee & Penrod, 2019) and pristine ID conditions (e.g., Table 1). The nominal size correction could also be applicable if filler selection is tailored to the reason of suspicion, such as if an innocent suspect is only in the lineup because they match an eyewitness description of the culprit and the fillers are also only in the lineup because they match that description. This scenario could also be simulated by the effective size correction, except that the conditions in the experiment would actually have to be pristine. This is because the effective size correction takes a measure of the number of plausible lineup members and assumes the innocent suspect is one of them. If the culprit-absent lineup in an experiment is constructed perfectly, resulting in an even distribution of IDs across the lineup, then the effective size correction would give estimates that are applicable to the same ideal case scenarios as the nominal size correction. Conversely, if some lineup members in the experiment are implausible and rarely chosen, estimates from the effective size correction would be applicable to investigative scenarios that have a comparable number of plausible lineup members and some reason that causes the innocent suspect to be among the plausible options. For instance, Steblay and Wells (2020) found that lineups in criminal cases often contain suspects who match the eyewitness description of the culprit better than do the fillers. Therefore, if the effective size of a lineup in an experiment is equal to three, the effective size correction could be used to estimate the risk to an

innocent suspect who matches an eyewitness description and appears in a lineup with only two fillers who also match the description. Given that neither of the corrections are applicable to all investigative scenarios, in the empirical work that follows we report both corrections (and no correction) to estimate the full PPV range.

Re-Analysis of Previous Studies

To compare the different methods of estimating the risk to innocent suspects, we obtained datasets from 10 eyewitness ID studies published by other authors. For a study to be included, it needed to report the distribution of mistaken IDs in the culprit-absent lineup so that we could compute its effective size. In the eyewitness literature researchers typically only report one false positive error rate that aggregates the mistaken IDs of all lineup members, which precluded us from conducting a comprehensive meta-analysis of the wider eyewitness literature. Instead, we obtained a convenience sample of openly available datasets that specified the distribution of eyewitness IDs across the culprit-absent lineup. After using these distributions to measure each lineup's effective size, we estimated PPV using the effective size correction, the nominal size correction, and no correction.

Method

The search procedure is reported in Panel A of Figure 2. Given that the distribution of lineup choices is not conventionally reported in published articles, we searched the Open Science Framework (OSF) for datasets that included the lineup choice distributions using "lineup" and "eyewitness identification" as keywords. We looked for experiments that presented a target during an encoding event (e.g., a crime video) and subsequently presented the target in a simultaneous photo lineup with a minimum of four lineup members. We excluded data from lineups that were intentionally biased against the suspect. For example, in a subgroup of one

study that otherwise met the inclusion criteria (Colloff et al., 2016), the suspect was the only lineup member with a black eye and that subgroup was excluded from the analysis. The search completed in December 2021 and yielded a final sample of 10 studies. Articles with included studies are marked with an asterisk in the references. Although unpublished studies with data on the OSF would have been eligible, the included studies were all published.

Study characteristics are reported in Table 3. All studies included culprit-present and culprit-absent conditions, with samples ranging from $N = 555$ to $N = 10{,}559$. In five of the studies, all participants within a subgroup completed the same lineup. In the other five studies, lineup members were sampled and rotated from a larger pool. For studies that used filler rotation we weighted each lineup member's ID probability by the number of times that the filler appeared in a lineup (for calculations, see Weighting in Filler Rotation Studies in online Supplemental Materials). Confidence ratings were used to further partition the data into three confidence subgroups: low = 0-59%, moderate = 60-89%, and high = 90-100%. Most studies used a 0-100% scale for confidence ratings, but one used a 5-point scale (Winsor et al., 2021), which we transformed to a 0-100% scale.

Lineups included between four and nine lineup members. To compute a summary estimate of the effective sizes in the studies, we adjusted the effective sizes of any lineups that deviated from six lineup members such that they would be equivalent to a 6-member lineup. For example, if nominal size = 8 and effective size = 4, then the effective size would be adjusted to 3 (4*[6/8] = 3) to make it comparable with 6-member lineups. This adjustment was only used for computing the summary estimate of effective size. The unadjusted nominal/effective size values were used in all computations of PPV.

PPV was estimated from guilty suspect IDs and innocent suspect IDs estimated via nominal size correction, effective size correction, or no correction. For the nominal size correction, the overall rate of mistaken IDs in the culprit-absent lineup was divided by the number of lineup members. For the effective size correction, the same overall ID rate was divided by the effective size of the lineup. Effective size was measured from the distribution of eyewitness choices in the culprit-absent lineups, using the formulas reported by Tredoux (1998). To account for variance in the plausibility of lineup members at different confidence levels, we used the lineup choice distributions in each confidence group to generate separate measures of effective size for low, moderate, and high confidence IDs. For no correction, every mistaken ID in the culprit-absent lineup was classified as an innocent suspect ID.

We used multilevel linear modeling in the metafor package in *R* (R Core Team, 2022; Viechtbauer, 2010) to generate summary estimates of effective size and PPV. Each study had subgroups that were not independent of one another, and Colloff and Wixted (2020) used stimuli that were used previously by Colloff et al. (2016). Therefore, we used the rma.mv function to compute a four-level hierarchical random effects model with effective size or PPV at Level 1, stimulus set groupings at Level 2, study groupings at Level 3, and author groupings at Level 4.

Availability of Data

The data are available at https://osf.io/b7gsu/.

Results

Effective lineup size. The average effective size for the lineups was 3.78, 95% CI [2.20, 5.36], $z = 4.70$, $p < .001$. Effective size was also moderated by eyewitness confidence, $Q(2) = 272.19$, $p < .001$. Choices were most evenly distributed among eyewitnesses who reported low confidence in their ID, $M = 4.02$, 95% CI [3.85, 4.18], and became increasingly less evenly

distributed as confidence increased; moderate confidence: M = 3.57, 95% CI [3.41, 3.73], high confidence: M = 2.67, 95% CI [1.81, 3.53].

PPV. Panel B of Figure 2 depicts the aggregate PPV range in the 10 studies. PPV scores have a range of 0 to 1. If PPV = 1, then all identified suspects are estimated to be guilty. If PPV = 0, then all identified suspects are estimated to be innocent. When confidence was 90% or greater, the PPV of suspect IDs was estimated to be 0.86 (95% CI [0.80, 0.93]) with the nominal size correction, 0.71 (95% CI [0.62, 0.79]) with the effective size correction, and 0.58 (95% CI [0.49, 0.67]) with no correction. The PPV range for each subgroup in the 10 experiments is plotted in Figure 3.

Discussion

Most lineups in the 10 previously published studies included a combination of plausible and implausible lineup members. Each lineup had, or was adjusted to have, six lineup members. If the lineups were perfectly fair, they would have had an effective size of six. The average effective size in the culprit-absent lineups, however, was less than four. In other words, based on the distribution of lineup choices, over one-third of the culprit-absent lineup members were classified as implausible options.

There was a wide range of effective size scores, leading to varying discrepancies between the nominal and effective size corrections. The discrepancy was most pronounced in two studies that rotated fillers from the larger pools (Akan et al., 2021; Colloff et al., 2016). In these studies, fillers were randomly selected from a pool that contained dozens of options. Finding a large pool of plausible fillers would normally be harder than finding five plausible fillers for a non-rotated lineup. Therefore, all else equal, rotation studies would be expected to have a disproportionate number of implausible fillers.

Regardless of which correction was applied, the suspect ID error rate in our re-analysis of 10 studies was higher than the 3% error rate estimated for the 15 studies reviewed by Wixted and Wells (2017). The suspect ID error rate is simply the inverse of PPV. In our sample, the error rate was 14% with the nominal size correction and 29% with the effective size correction. The effective size correction accounts for implausible lineup members, so a higher error rate with this approach was to be expected. It is not entirely clear why we observed a higher error rate with the nominal size correction than that observed by Wixted and Wells, given that they also applied this correction. One possible explanation for the discrepancy is that the studies we found on the OSF were all recent publications, and there is no overlap between the studies aggregated by us and the studies aggregated by Wixted and Wells. Another consideration is that Wixted and Wells included only studies categorized as pristine. Other than excluding conditions that were intentionally biased against an innocent suspect, pristine conditions were not part of our inclusion criteria.

Limitations. Just as the estimates from Wixted and Wells (2017) are meant only to be generalized to eyewitness ID under pristine conditions, there are several constraints on generality to note for the sample of datasets we analyzed (Simons et al., 2017). The studies we reviewed all had limits to ecological validity, which inevitably creates doubt as to whether the estimates in our analysis are representative of error rates in real criminal cases. This was also not a comprehensive review of the eyewitness literature, and we instead limited our meta-analysis to openly-available datasets that could be easily found on the OSF. There are many characteristics of the sample that were advantageous, such as the large sample sizes and wide range of lineups that were tested (Wells & Windschitl, 1999). Nonetheless, most eyewitness researchers do not post their data on the OSF, and when they do the lineup choice distribution is not always

included. Many of the studies we reviewed were also from the same author group. Therefore, the estimates are unlikely to be representative of the wider literature.

Experiments 1—3

We conducted three experiments with lineups that were fair or biased against a designated innocent suspect. Matching lineup fillers to features previously described by the eyewitness is crucial for creating a fair lineup (Luus & Wells, 1991; Wells et al., 2020). If lineup fillers do not match the eyewitness description, an innocent suspect that does match it faces an increased risk of misidentification (Wells et al., 1993). In our experiments, we constructed lineups with fillers whose hair was similar or dissimilar to the appearance of the designated innocent suspect. This type of filler bias is known to increase mistaken IDs of innocent suspects (Clark, 2012; Fitzgerald et al., 2013; Lindsay & Wells, 1980; Smith et al., 2017).

Hypotheses

Prior to collecting data for Experiment 3, we preregistered the following hypotheses (https://aspredicted.org/KSG_54R): (1) The observed rate of designated innocent suspect IDs will be higher in biased lineups than in fair lineups; (2) When the effective size correction is used, the estimated rate of innocent suspect IDs will be higher in biased lineups than in fair lineups; (3) When the nominal size correction is used, the estimated rate of innocent suspect IDs will not significantly differ between biased and fair lineups. Experiment 1 and 2 were designed for a different purpose, and no hypotheses involving the method of classifying innocent suspect IDs were preregistered for those experiments.

Method

General design. In each experiment we manipulated the presence of the culprit in the lineup and the fairness of the lineup fillers in relation to a suspect (Figure 4). Culprit presence

was manipulated using the single lineup paradigm (Oriet & Fitzgerald, 2018). In this paradigm, participants are assigned to witness crimes with different culprits and then assigned to the same lineup, which contains one of the culprits. The benefit of this method is that the culprit-present and culprit-absent lineups are identical, which allows the fillers to always be matched to the appearance of the suspect. Lineup bias was manipulated by selecting fillers who varied in their match to the suspect's hair length or hair color. Although none of the lineups was perfectly fair, for brevity we refer to the lineups with matched hair as "fair lineups" and the ones with mismatched hair as "biased lineups."

General procedure. All experiments were advertised as eyewitness ID studies, and all participants completed the self-administered experiments online. Participants watched a staged crime video, completed a 3-minute computer game (*Snakes*, e.g., playsnake.org) as a filler task, and then completed a single lineup ID task. Participants were instructed to watch the crime video in a place without distractions and to pay attention because they would be asked about their perceptions of the video afterwards. The crime videos showed a woman entering a hallway or room, looking through an unattended bag, and stealing a laptop from the bag. At the end of the video, an unrelated image appeared on the screen (e.g., a banana). Participants were then asked to identify the image via a 6-item multiple choice attention check question, and anyone who selected an image other than the one presented were excluded. After the manipulation check participants were informed that a lineup would be presented and that the thief from the video may or may not be present. After the pre-lineup warning, a simultaneous photo lineup appeared with instructions to select the person from the video if she was present or to select the "Not Present" option if absent. After providing a lineup decision, participants rated their confidence from 0 (not sure) to 100 (sure), answered questions about the reasons for their ID decision, and

were thanked for participating. The protocols for all experiments were reviewed by an institutional review board.

Experiment 1. Participants were recruited from the Qualtrics survey panel (www.qualtrics.com), with the intention to collect at least 100 participants (after exclusions) in each condition of a 2 (culprit-present vs. culprit-absent) × 2 (fair lineup vs. biased lineup) design. This sample size gives > .90 power to detect medium-sized differences (Cohen's $h = 0.50$) in ID rates. The final sample was $N = 686$. For further details, see Participant Demographics and Participant Exclusions in online Supplemental Materials.

Participants were randomly assigned to watch a crime video with Culprit A or Culprit B. The actor who played Culprit A was the designated suspect in all lineups, whereas Culprit B did not appear in any lineups. Accordingly, the lineup was culprit-present for participants who witnessed Culprit A and culprit-absent for participants who witnessed Culprit B. All lineups contained one suspect and five fillers, with position of the suspect in the lineup counterbalanced across participants. Both culprits were Black women with long braided hair (Figure 4). In the fair lineup, the fillers were also Black women with long braided hair. In the biased lineup, the fillers were Black women, but their hair was short and/or unbraided. We initially planned to use a different person as Culprit B, but pilot testing indicated participants were unable to discriminate between that person and Culprit A (see Pilot Study 1 in online Supplemental Materials). We assigned a disproportionate number of participants to the culprit-present condition (see Experiment 1 Programming Error in online Supplemental Materials), so some analyses required a correction to equate the culprit-present and culprit-absent sample sizes (see Analysis).

The experiment was as described in the General Procedure except that half of the participants were randomly assigned to provide similarity ratings before they completed the

lineup. These participants viewed the lineup and rated each lineup member's similarity to their memory of the culprit on a scale from 0% (not similar) to 100% (similar). A 2 (fair vs biased) × 2 (innocent suspect vs fillers) mixed ANOVA on culprit-absent lineup data yielded a main effect of lineup member, such that similarity ratings were higher for the innocent suspect than for the fillers, M = 51.1 vs. M = 27.6, respectively, $F(1,114) = 62.55$, $p < .001$, d = 0.68, 95% CI [0.47, 0.88]). Accordingly, we classified this experiment to have a best-match innocent suspect. The ANOVA also yielded a significant interaction, $F(1,114) = 13.28$, $p < .001$, $\eta_p^2 = .10$, which indicated that although the similarity ratings were higher for the innocent suspect than for the fillers irrespective of lineup bias, the difference was more pronounced in the biased lineup (innocent suspect = 59.4 vs fillers = 24.0, d = 1.12, 95% CI [0.77, 1.45]) than in the fair lineup (innocent suspect = 43.9 vs fillers = 30.8, d = 0.39, 95% CI [0.13, 0.64]). For additional data, see Similarity Ratings in online Supplemental Materials.

Experiment 2. Participants were recruited from the Qualtrics panel, with the intention to collect at least 100 participants (after exclusions) in each condition of the 2 (culprit-present vs. culprit-absent) × 2 (fair lineup vs. biased lineup) design. The final sample included for analysis in the present research was N = 405. For further details, see Participant Demographics and Participant Exclusions in online Supplemental Materials. Hypotheses were pre-registered but they are not relevant for comparing estimation methods.

The method was nearly identical to Experiment 1. One exception is that the lineups included 4, 6, or 8 lineup members. In the 4- and 6-member lineups the fillers were sampled from the pool of seven fillers. To circumvent the complications of computing the effective size of lineups with rotated fillers, we only analyzed the 8-member lineups from this experiment. Across participants, the suspect appeared in Position 1, 3, 5, or 7.

A second exception is that all participants provided similarity ratings, and the ratings were always made after making the categorical lineup decision. A 2 (fair vs biased) × 2 (innocent suspect vs fillers) mixed ANOVA on culprit-absent lineup data showed that, again, the innocent suspect was rated as more similar than the fillers to the participants' memories of the culprit, M = 49.6 vs M = 22.8, respectively, $F(1,180) = 123.29$, $p < .001$, $d = 0.75$, 95% CI [0.59, 0.91]. The interaction was also significant: Similarity ratings were higher for the innocent suspect than for the fillers irrespective of lineup bias but the difference was more pronounced in the biased lineup (innocent suspect = 60.0 vs fillers = 17.2, $d = 1.18$, 95% CI [0.92, 1.44]) than in the fair lineup (innocent suspect = 38.4 vs fillers = 28.7, $d = 0.37$, 95% CI [0.16, 0.59]), $F(1,180) = 49.08$, $p < .001$, $\eta_p^2 = .21$. For additional data, see Similarity Ratings in online Supplemental Materials.

Experiment 3. Participants in Experiment 3 were recruited from MTurk using Cloud Research's MTurk Toolkit (Litman et al., 2017), with eligibility limited to those with experience of 500+ approved HITs and an approval rate of 95% or higher. Our aim was to collect data from 1500 participants in a 2 (culprit-present vs. culprit-absent) × 2 (fair lineup vs. biased lineup) design. In culprit-absent lineups, we also manipulated the similarity between the perpetrator at the crime and the innocent suspect in the lineup (best match vs next best match vs weak match). The final sample had N = 1531. For further details, see Participant Demographics and Participant Exclusions in online Supplemental Materials. This sample size gives >.95 power to detect medium-sized differences (Cohen's $h = 0.50$) in ID rates.

A new stimulus set was used in Experiment 3. This time the culprits and lineup members were White women, and the innocent suspects stood out from the fillers in biased lineups because of their hair color rather than hair length. Participants watched a crime video with Culprit C or Culprit D, both of whom had light brown hair. Fair lineup fillers also had light

brown hair, whereas biased lineup fillers had blonde hair. All lineups contained six people, and position of the suspect in the lineup was randomly determined. Pilot research showed that the fair lineups included light-brown-haired lineup members who were frequently confused with culprits and biased lineups included blonde lineup members who were infrequently confused with the culprit. For further details, see Pilot Study 2 in online Supplemental Materials.

In addition to using a new stimulus set, Experiment 3 differed from the first two experiments in several respects. First, we counterbalanced whether Culprit C or Culprit D was the suspect in the lineup. Culprit C and Culprit D were not commonly mistaken for each other in pilot research, so we classified these two actors as weak-match innocent suspects in each other's culprit-absent lineups (Figure 4). Second, three biased culprit-absent lineups were tested for Culprit D. In addition to the biased lineup containing a weak match, with Culprit D as the innocent suspect, we tested a second biased lineup with a best-match innocent suspect and a third lineup with a "next-best" innocent suspect. The best-match innocent suspect was the light-brown-haired lineup member most frequently mistaken for Culprit D in the pilot study. We had planned to do similarly for Culprit C; however, due to a programming error the video for Culprit D was presented to participants assigned to the biased lineup with the best-match for Culprit C, resulting in a total of three biased lineups for Culprit D and only one biased lineup for Culprit C (Figure 4). The third innocent suspect for Culprit D was the second most likely lineup member to be mistaken for Culprit D in the pilot research, so we classified her as the next-best innocent suspect, borrowing the terminology of Clark and Davey (2005). Third, after the filler task and before the pre-lineup instructions, participants described the person from the video on the following features: sex, race/ethnicity, build, age, hair, and distinctive features. Note that no similarity ratings were collected in this experiment.

Analysis. Before computing PPV, we corrected the data for unequal sample sizes in the culprit-present and culprit-absent lineups. If the number of participants in the culprit-present and culprit-absent lineups is equal, then the chance likelihood that an identified suspect would be guilty is 50%. To account for different sample sizes across culprit-present and culprit-absent conditions, we divided the culprit-absent *n* by the culprit-present *n* and multiplied the guilty suspect ID frequency by this correction term (and then rounded to the nearest whole number if necessary). This adjusted the chance likelihood of suspect guilt to 50% (i.e., a 50% base rate).

PPV was calculated by dividing guilty suspect IDs by the total number of suspect IDs (guilty + innocent). For innocent suspect IDs, we used either IDs of the designated innocent suspect or an estimate derived from the nominal size correction, the effective size correction, or no correction. After grouping participants by confidence, two types of effective size corrections are possible. In the *variable effective size correction*, which was applied in the re-analysis of the 10 previously published studies, a separate estimate of the lineup's effective size was generated from the lineup choice distributions in each confidence group and used to estimate the innocent suspect ID rate for the associated level of confidence. We also applied a *fixed effective size correction*, which used a single measure of effective size generated from the overall distribution of lineup choices (i.e., before splitting into confidence groups) and applied this constant measure of effective size to estimate innocent suspect ID rates for each of the confidence groups. The fixed effective size correction accounts for the overall plausibility of the lineup members, whereas the variable effective size correction additionally accounts for variance in the plausibility of lineup members at different confidence levels.

We used the Metafor package in R to compute summary effect sizes using a random effects model. Odds ratios are reported as the effect size for differences in ID responses, with

95% confidence intervals reported in square brackets. Cochrane's *Q* is reported as a significance test for effect size heterogeneity.

Availability of stimuli and data. Stimuli and data are available at https://osf.io/b7gsu/.

Results

Manipulation check. Effective lineup size was higher in fair lineups than in biased lineups for all stimulus sets (Table 4), confirming that the filler manipulation produced the intended effect on lineup bias.

Suspect IDs. Lineup bias increased correct IDs. Figure 5 shows that in culprit-present lineups, the odds of a guilty suspect ID were on average more than four times as great in biased lineups than in fair lineups, $OR = 4.26$ [2.38, 7.62], $z = 4.88$, $p < .001$. Significant effect size heterogeneity was detected (see *Q* test in Figure 5), with larger effect sizes in Experiment 3 than in the first two experiments. Lineup bias had minimal effect on overall choosing from culprit-present lineups, and thus the decrease in correct IDs in fair lineups coincided with an increase in filler IDs (see Table 5).

Lineup bias also increased IDs of the designated innocent suspect in culprit-absent lineups, and the effective size correction was the only method of estimation that detected this effect of lineup bias (see Figure 6). The odds of misidentifying the designated innocent suspect were on average more than five times as great in biased versus fair lineups, $OR = 5.50$ [2.77, 10.95], $z = 4.94$, $p < .001$, with significant effect size heterogeneity due to larger effects of bias when the innocent suspect was a weak match, $Q(5) = 24.45$, $p < .001$. When the effective size correction was applied, the estimated rate of innocent suspect IDs was also higher in biased versus fair lineups, $OR = 3.04$ [2.13, 4.33], $z = 6.43$, $p < .001$, and effect size heterogeneity did not reach the significance threshold, $Q(5) = 10.79$, $p = .056$. By contrast, when the nominal size

correction was applied, no effect of lineup bias on innocent suspect IDs was detected, *OR* = 0.84 [0.60, 1.18], *z* = 0.98, *p* = .329, with minimal effect size heterogeneity, *Q*(5) = 1.27, *p* = .938. If no correction was applied and the overall rate of misidentifications was used, the direction of the effect on designated innocent suspect IDs was reversed: increasing lineup bias was estimated to reduce the odds of misidentification, *OR* = 0.72 [0.52, 0.98], *z* = 2.12, *p* = .034, with significant effect size heterogeneity, *Q*(5) = 12.14, *p* = .033.

PPV. The effect of lineup bias on the PPV of suspect IDs is depicted in Figure 7. If calculated using designated innocent suspect IDs as the false alarm rate, PPV was not significantly affected by lineup fairness, *OR* = 0.66 [0.36, 1.19], *z* = 1.40, *p* = .163; however, significant heterogeneity in effect sizes was present, *Q*(5) = 12.14, *p* = .033, and increasing lineup bias led to significantly lower PPVs in both tests involving a weak match innocent suspect. When the effective size correction was used, again no effect of lineup fairness on PPV was detected, *OR* = 1.03 [0.63, 1.69], *z* = 0.13, *p* = .900; for this analysis, significant heterogeneity in effect sizes was present, *Q*(5) = 17.39, *p* = .004.

When the nominal size correction was applied, increasing lineup bias was estimated to increase PPV, *OR* = 2.79 [1.78, 4.36], *z* = 4.48, *p* < .001. No significant heterogeneity was detected, *Q*(5) = 7.40, *p* = .193. If no correction was applied, again lineup bias was estimated to increase PPV, *OR* = 2.77 [1.80, 4.27], *z* = 4.63, *p* < .001. Significant heterogeneity was detected with no correction, *Q*(5) = 23.95, *p* = .193, which is a consequence of the larger sample size when no correction was applied compared to the nominal size correction, which included only 1/6 of the mistaken IDs from the culprit-absent lineup.

Eyewitness Confidence and Lineup Choice Distributions. Figures 8 and 9 depict the distributions of IDs across the members of culprit-present and culprit-absent lineups,

respectively. Data are reported separately for eyewitnesses with low (0-59%), moderate (60-89%), and high confidence (90-100%). Effective size values for culprit-absent lineups are also reported in Figure 9. The idea of calculating more than one effective size estimate for the same lineup may seem counterintuitive. Effective size is generally interpreted as a measure of lineup fairness, and perhaps lineup fairness should not depend on whether a witness is 50% or 100% confident. But even if lineup fairness cannot change with eyewitness confidence, Figure 9 suggests the proportion of IDs that a lineup member attracts could change with confidence. For example, in Experiment 1, Filler #2 attracted 22% of the IDs made with low confidence, 7% of the IDs made with medium confidence, and 0% of the IDs with high confidence. This suggests Filler #2 was plausible enough to be misidentified by an uncertain witness, but not plausible enough for a confident eyewitness to misidentify them. Thus, rather than variations in lineup fairness across confidence groups, the effective size (E) values in Figure 9 reflect the number of lineup members who were plausible enough to be identified at each level of confidence.

PPV range. We computed summary estimates of PPV for participants with low (0-59%), moderate (60-89%), and high confidence (90-100%). Each summary estimate represents a weighted average of four PPVs, one for each of Experiments 1 and 2, and one for each of Culprits C and D in Experiment 3. For Culprit D, there was one frequency of guilty suspect IDs and three corresponding frequencies of innocent suspect IDs (best, next-best and weak), so we computed three PPV estimates (i.e., using each of the three innocent suspect ID frequencies) and used the average of these PPV estimates for Culprit D.

Figure 10 depicts the PPV range for fair and biased lineups. In fair lineups (Panel A), no correction resulted in a noticeably lower estimate of PPV than the other four curves. For biased lineups (Panel B), applying the nominal size correction to the innocent suspect ID rate resulted in

a PPV curve that was noticeably higher than the other four curves. Error rates for high confidence suspect IDs from fair lineups ranged from .06 to .37, depending on how innocent suspect IDs were classified. Of the high confidence IDs from fair lineups, 90.6% were guilty suspects and 9.4% were designated innocent suspects. In other words, the designation approach resulted in a suspect ID error rate of .09 [.01, .21]. Estimating innocent suspect IDs in fair lineups yielded corresponding error rates of .06 [.00, .17] with the nominal size correction, .10 [.01, .21] with the fixed effective size correction, .17 [.06, .32] with the variable effective size correction, and .37 [.23, .53] with no correction.

Error rates for high confidence suspect IDs from biased lineups were notably higher than for fair lineups, except when estimated using the nominal size correction. When designated innocent suspect IDs were used, the error rate for high confidence suspect IDs in biased lineups was .25 [.17, .33]. This corresponds with estimated error rates for biased lineups of .19 [.13, .27] using the variable effective size correction, .23 [.16, .31] using the fixed effective size correction, .27 [.19, .36] with no correction, and only .05 [.01, .11] using the nominal size correction.

Discussion

The interpretation of lineup bias effects depended on how innocent suspect IDs were classified. As predicted in Hypothesis 1, the designated innocent suspect ID rate was higher in biased lineups than in fair lineups. The implausible fillers in biased lineups also made it easier to correctly identify the guilty suspect in culprit-present lineups, so lineup bias had no effect on PPV if designated innocent suspect IDs were used as the false alarm rate. Consistent with Hypothesis 2, increasing lineup bias also increased the estimate of innocent suspect IDs with the effective size correction, which again negated any benefit to PPV from the gain in guilty suspect IDs in biased lineups. By contrast, the nominal size correction resulted in higher PPV estimates

in biased lineups than in fair lineups. This is because lineup bias increased IDs of guilty suspects and, as predicted in Hypothesis 3, lineup bias had no effect on the nominal-size-corrected estimate of innocent suspect IDs. In effect, the nominal size correction statistically eliminates the cost of bias in culprit-absent lineups but makes no corresponding adjustment to correct for the benefit of bias in culprit-present lineups. This has clear implications for estimating PPV with the nominal size correction, which we consider further in the General Discussion.

For high confidence suspect IDs, again the error rate depended on how innocent suspect IDs were classified. With the nominal size correction, approximately 5% of suspect IDs in fair lineups were estimated to be errors. This is consistent with Wixted's (2018) estimate that approximately 95% of suspect IDs would be accurate under pristine conditions. In our experiments, however, the suspect ID error rate in fair lineups was approximately twice as high if measured from designated innocent suspect IDs (9%) or if estimated with the fixed effective size correction (10%). This is because even though we aimed to make these lineups fair, the distribution of IDs across the lineup members shows that they were not perfectly fair (Figure 9) and, contrary to the nominal size correction, these alternative methods of classifying innocent suspects account for lineup bias and assume the bias would affect the risk to innocent suspects.

Limitations. Like most research on eyewitness ID, the generalizability of our findings is limited by unrealistic witnessing and testing conditions. The experiments were completed online in a single session lasting 5-10 minutes. Participants also knew the crime they witnessed was staged, which can affect lineup decisions (Eisen et al., 2022). Although such experiments are useful for understanding relative differences in the risk to an innocent suspect (Wells & Quinlivin, 2009), greater ecological validity would be necessary before extrapolating the precise error rates observed to real criminal cases. An additional limitation is that sample sizes became

quite small when focusing on high confidence IDs, especially for fair lineups. This has particular implications for the precision of estimates from the variable effective size correction, which relied on effective size estimates from small samples of high confidence IDs. One further potential limitation is that all estimation methods assume that lineup bias does not impact the overall rate of choosing from culprit-absent lineups. Although the lineup rejection rates in Table 5 did not consistently show an effect of lineup bias on choosing in our experiments, lineup bias did reduce the choosing rate for Culprit C. If lineup bias influences eyewitnesses' inclination to choose from the lineup, it would affect the precision of any estimation method involving a correction to the total number of mistaken IDs.

General Discussion

Not all methods of estimating innocent suspect IDs are sensitive to lineup bias. Across three experiments, the nominal size correction indicated that implausible lineup members had no effect on the risk of innocent suspect IDs. Applying no correction was similarly insensitive to lineup bias and only approximated IDs of the designated innocent suspect if the lineup was biased. Rather than adjusting for lineup bias, the nominal size correction merely shows the risk to innocent suspects *if* the lineups had been fair and applying no correction only shows the risk *if* all the ID errors were innocent suspect IDs. By contrast, the effective size correction adjusts the overall mistaken ID rate using an empirical measure of lineup fairness. If the mistaken IDs are not equally distributed across the lineup and instead concentrate on certain lineup members, the effective size correction assumes the innocent suspect is one of the more plausible lineup members and estimates an innocent suspect ID rate that is invariably higher than the estimate from the nominal size correction and invariably lower than the estimate from no correction.

How to Classify Innocent Suspect IDs in Lineup Experiments

Designation or estimation can be used in experiments to classify lineup errors as innocent suspect IDs or filler IDs. In our three experiments, designation gave the most informative measure of the innocent suspect ID rate. This is because we manipulated how much the lineup fillers matched the appearance of a designated innocent suspect, and filler selection is a suspect bias variable (Wells & Olson, 2001). Designation would be harder to justify if we had manipulated a general impairment variable. For example, if our only manipulation had been the retention interval between witnessing the event and completing the lineup, we would have had no reason for designating one lineup member as the innocent suspect over any of the others. In such circumstances, estimation can be used to avoid having to arbitrarily designate one lineup member to be the innocent suspect.

Estimates from the nominal size correction are most applicable to innocent suspects who appear in lineups under pristine conditions. Steblay and Wells (2020) assessed lineup fairness in real criminal cases and found that, across samples from several jurisdictions, suspects were often the best match to an eyewitness description of the culprit. In one analysis, Steblay and Wells tested 190 lineups from a field experiment (Wells et al., 2015) and found that 43% were fair, 33% were suspect-biased, and 24% were reversed-biased (i.e., the suspect was less likely than fillers to best match the description). The nominal size correction assumes that an innocent suspect would be no more plausible than the average filler. Therefore, assuming that eyewitnesses focus on lineup members that match their prior description, the nominal size correction would underestimate the risk to innocent suspects in the suspect-biased lineups. Under the same logic, the nominal size correction would be expected to effectively estimate the risk in the fair lineups (and overestimate the risk in the reverse-biased lineups). Note, however, that

only one aspect of bias was measured by Steblay and Wells. The nominal size correction would only apply in the absence of any other type of suspect bias (for a list of 10 suspect variables, see Smalarz, 2021).

Although it is not always possible to know if suspect bias is present (Sauer et al., 2019), the reason of suspicion in the case is one suspect bias variable that would often be known. The nominal size correction assumes the reason of suspicion in the case has no effect on the innocent suspect's risk, but some reasons of suspicion clearly do increase the risk to innocent suspects. This is demonstrated in the DNA exoneration case of Thomas Haynesworth, who became an innocent suspect when a victim mistook him for the perpetrator during a chance encounter on the street (*Thomas Haynesworth vs. Commonwealth of Virginia*, 2011). An innocent suspect is unlikely to be spontaneously misidentified on the street unless they resemble the actual culprit, and a strong resemblance between Haynesworth and the culprit could also explain how four additional victims ended up misidentifying Haynesworth at police lineups. Lee and Penrod (2019) found that innocent suspect ID rates were higher if an innocent suspect was designated than if they were estimated with the nominal size correction. Accordingly, Lee and Penrod proposed that designation was more representative of cases with appearance-based reasons of suspicion and that the nominal size correction was more representative of cases with non-appearance-based reasons of suspicion.

The effective size correction is a method of estimating the risk to innocent suspects under non-pristine conditions. In this method the number of plausible lineup members is measured, and the innocent suspect is presumed to be one of them. A variety of investigative factors could increase the plausibility of an innocent suspect. For example, if the suspect is under investigation because they match the eyewitness description of the culprit and the lineup fillers are not equally

matched to that description, then the innocent suspect would probably be one of the more plausible options to the eyewitness. This type of bias does not appear to be especially rare. Over one-third of the lineups analyzed by Steblay and Wells (2020) included suspects who best matched the eyewitness description more often than expected by chance.

Other criminal case scenarios would be less well represented by the effective size correction. Matching the eyewitness description of the perpetrator is only one of several reasons of suspicion that increase the risk of innocent suspect IDs (Wells & Penrod, 2011). If an innocent suspect was under investigation because their appearance closely matched that of a culprit caught committing a crime on a CCTV image, conventional methods of lineup construction such as matching fillers to the eyewitness description would be unlikely to result in other lineup members who are as plausible as the innocent suspect. In this type of scenario, the innocent suspect would not only be one of the plausible lineup members – they would be the most plausible lineup member, which has led to suggestions that a lineup should not be conducted at all when the reason of suspicion is so likely to result in suspect bias (Shen et al., 2023). In such scenarios, the effective size correction would underestimate the risk to innocent suspects because it only assumes that the innocent suspect would be one of the more plausible lineup members, not the most plausible one. Accordingly, this lineup scenario would be better represented by using a designated innocent suspect who closely matches the appearance of the culprit (Lee & Penrod, 2019).

Given that no correction is universally applicable to all eyewitness scenarios, reporting only the nominal size correction could give a distorted impression of the risk to innocent suspects. The nominal size correction assumes being a suspect has no effect on an innocent lineup member's risk of being mistakenly identified. This assumption would only be valid if

suspect bias was entirely absent (Hyman, 2021). Reasonable people can disagree on how often this happens in practice, but relying solely on the nominal size correction would only be justified if eyewitness ID conditions were always pristine or if the research question was entirely focused on error rates under pristine conditions. Ultimately, the method of classifying innocent suspect IDs should depend on the objectives of the research. In most cases, a designated innocent suspect would be preferred for studying suspect bias variables and estimation would be preferred for studying general impairment variables. If estimation is used, we recommend reporting the full PPV range, as depicted in Figure 2, and emphasizing that each estimate is contingent upon the underlying assumptions and applies to different ID scenarios.

How to Estimate the Suspect ID Error Rate

The suspect ID error rate is the inverse of PPV, and our experiments show that the nominal size correction gives a distorted impression of PPV if lineups are not perfectly fair. If a witness has previously reported that the culprit has curly hair and the only lineup member with curly hair is the suspect, this bias would be expected to increase suspect IDs regardless of whether the suspect is guilty or innocent (Lindsay & Wells, 1980; Wells et al., 1993; Clark, 2012). With the designation approach, this is precisely what we found: lineup bias increased IDs of both the guilty suspect and the designated innocent suspect. The nominal size correction, however, creates an analytic asymmetry between guilty and innocent suspects: It eliminates the effect of bias against innocent suspects by imposing an artificial cap on the innocent suspect ID rate, but it makes no adjustment to the guilty suspect ID rate in the culprit-present lineup (Quigley-McBride & Wells, 2021). This is a problem because correct IDs in a biased lineup could be lucky guesses, enabled by the inclusion of implausible fillers (Wells et al., 2012). Thus, if the same implausible fillers appear in both the culprit-present and culprit-absent lineups, the

bias that should be increasing guilty and innocent suspect IDs would only be realized in the measure of guilty suspect IDs, resulting in a deceptively high ratio of guilty-to-innocent suspect IDs. This explains why using the nominal size correction in our experiments resulted in a lower suspect ID error rate in biased lineups than in fair lineups.

This finding gives reason to be skeptical of nominal-size-corrected estimates of the suspect ID error rate. The nominal size correction imposes a ceiling on the innocent suspect ID rate to 1/*k*, where *k* is nominal lineup size. Assuming that a lineup has six members and no correction is applied to the guilty suspect ID rate, biased fillers would have the potential to increase the guilty suspect ID rate to 100% but at most could increase the innocent suspect ID rate to only 17% (Quigley-McBride & Wells, 2021). Because of this analytic asymmetry, any suspect bias variable could be used in combination with the nominal size correction to artificially deflate the suspect ID error rate. A better approach to estimating PPV is to classify innocent suspect IDs using either designation or the effective size correction, which do not constrain the ceiling on the innocent suspect ID rate.

Future Directions

One question for future research is how often an innocent suspect would be among the plausible lineup members. Steblay and Wells (2020) found that innocent suspects in criminal cases were generally more likely than the average filler to be the best match to the eyewitness description, but they also found that some innocent suspects were poor matches to the description. This shows that innocent suspects are not always plausible in relation to the eyewitness description. However, this is only one of many possible suspect bias variables (Smalarz, 2021). Understanding how often police lineups have any form of suspect bias would speak to how broadly the nominal and effective size corrections can be applied.

Our findings highlight the need for more research on the suspect ID error rate. One benefit of our meta-analysis over the previous analysis by Wixted and Wells (2017) is that we were able to estimate the effective size of the lineups and use it to estimate error rates on the assumption the suspect would be a plausible member of the lineup. However, our search for the meta-analysis was limited to publicly available datasets. For a more representative analysis of the literature in future research, researchers would need to start reporting how the mistaken IDs in their experiments are distributed across the lineup members (as in Figures 8 and 9), rather than just reporting the overall number of mistaken IDs. In addition to enabling computation of the effective size correction, reporting the number of times each lineup member was identified would increase transparency regarding the fairness of the lineups in the experiment.

Another consideration for future research is whether the fixed or variable effective size correction provides a better estimate of innocent suspect IDs. If the distribution of lineup choices is consistent across confidence groups, then both corrections should provide similar estimates but the fixed correction might be preferred nonetheless because it only requires one measure of lineup fairness. However, if the distribution of lineup choices is linked to ID confidence, as was found in our re-analysis of previous published literature, the variable effective size correction would account for this and may better reflect the risk to innocent suspects for each confidence group. The variable effective size estimate for the 90-100% group could also be more applicable to criminal cases, which are more likely to be prosecuted if the eyewitness is highly confident.

Conclusion

Wixted and Wells (2017) estimated that high confidence IDs under pristine experimental conditions resulted in a relatively low suspect ID error rate. This estimate, however, relied on the nominal size correction, which assumes an equal risk of misidentification for innocent suspects

and innocent fillers. Although we do not know if the lineups analyzed by Wixted and Wells were fair, in our re-analysis of 10 previous studies we found that lineups usually included a mix of plausible and implausible lineup members. In our experimental research, we found that when the nominal size correction is applied for lineups with implausible fillers, the fillers increased the guilty suspect ID rate and the nominal size correction limited the potential for increases in the innocent suspect ID rate, resulting in an artificially low suspect ID error rate. We therefore conclude that unless the fairness of the lineups has been demonstrated, caution is warranted when interpreting suspect ID error rates that have been estimated with the nominal size correction.

References[1]

*Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2021). The effect of lineup size on eyewitness identification. *Journal of Experimental Psychology: Applied, 27*(2), 369–392. https://doi.org/10.1037/xap0000340

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8,* 44-56. https://doi.org/10.1037//1076-898X.8.1.44

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

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

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior*, *29*(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*, 187-218. https://doi.org/10.1007/s10979-006-9082-4

[1] References marked with an asterisk indicate inclusion in our re-analysis of previous studies

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

*Colloff, M. F., & Wixted, J. T. (2020). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology: Applied, 26*(1), 124–143. https://doi.org/10.1037/xap0000218

*Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(8), e2017292118. https://doi.org/10.1073/pnas.2017292118

*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

Cutler, B. L., & Penrod, S. D. (1989). Moderators of the confidence-accuracy correlation in face recognition: The role of information processing and base-rates. *Applied Cognitive Psychology*, *3*, 95-107. https://doi.org/10.1002/acp.2350030202

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

Eisen, M. L., Ying, R. C., Chui, C., & Swaby, M. A. (2022). Comparing witness performance in the field versus the lab: How real-world conditions affect eyewitness decision-making. *Law and Human Behavior, 46*, 175–188. https://doi.org/10.1037/lhb0000485

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

Innocence Project (2023). Eyewitness misidentification. https://innocenceproject.org/eyewitness-misidentification/

Juncu, S., & Fitzgerald, R. J. (2021). A meta-analysis of lineup size effects on eyewitness identification. *Psychology, Public Policy, and Law, 27*(3), 295-315. https://doi.org/10.1037/law0000311

Kovera, M. B., & Evelo, A. J. (2017). The case for double-blind lineup administration. *Psychology, Public Policy, and Law, 23*, 421-437. https://doi.org/10.1037/law0000139

Kovera, M. B., & Evelo, A. J. (2021). Eyewitness identification in its social context. *Journal of Applied Research in Memory and Cognition 10, 313–327.* *https://doi.org/10.1016/j.jarmac.2021.04.003*

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

Lindsay, R. C., Wallbridge, H., & Drennan, D. (1987). Do the clothes make the man? An exploration of the effect of lineup attire on eyewitness identification accuracy. *Canadian Journal of Behavioural Science, 19*(4), 463–478. https://doi.org/10.1037/h0079998

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4,* 303-313. https://doi.org/10.1007/BF01040622

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556–564. https://doi.org/10.1037/0021-9010.70.3.556

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*, 433-442. https://link.springer.com/article/10.3758/s13428-016-0727-z

Luus, C. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15*, 43-57. https://doi.org/10.1007/BF01044829

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

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*, 93–102. https://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L. (2016). The effects of verbal descriptions on eyewitness memory: Implications for the real-world. *Journal of Applied Research in Memory and Cognition, 5*, 270-276. https://doi.org/10.1016/j.jarmac.2016.07.003

*Nyman, T. J., Lampinen, J. M., Antfolk, J., Korkman, J., & Santtila, P. (2019). The distance threshold of reliable eyewitness identification. *Law and Human Behavior, 43*(6), 527–541. https://doi.org/10.1037/lhb0000342

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law and Human Behavior*, 42, 1-12. http://dx.doi.org/10.1037/lhb0000272

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews, 10,* 1-11. https://doi.org/10.1186/s13643-021-01626-4

Pryke, S., Lindsay, R. C. L., Dysart, J. E., & Dupuis, P. (2004). Multiple independent identification decisions: A method of calibrating eyewitness identifications. *Journal of Applied Psychology, 89,* 73-84. https://doi.org/10.1037/0021-9010.89.1.73

Quigley-McBride, A., & Wells, G. L. (2021). Methodological considerations in eyewitness identification experiments. In A. M. Smith, M. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks* (pp. 85-112). Routledge.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law, 25*, 147–165. https://doi.org/10.1037/law0000203

*Seale-Carlisle, T. M., Colloff, M. F., Flowe, H. D., Wells, W., Wixted, J. T., & Mickes, L. (2019). Confidence and response time as indicators of eyewitness identification accuracy in the lab and in the real world. *Journal of Applied Research in Memory and Cognition, 8*(4), 420–428. https://doi.org/10.1016/j.jarmac.2019.09.003

Shen, K. J., Colloff, M. F., Vul, E., Wilson, B. M., & Wixted, J. T. (2023). Modeling face similarity in police lineups. *Psychological Review, 130*, 432-461. https://doi.org/10.1037/rev0000408

Simons, D. J., Shoda, Y., Lindsay, D. S. (2017) Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12,* 1123-1128. https://doi.org/10.1177/1745691617708630

Smalarz, L. (2021). Suspect bias: A neglected threat to the reliability of eyewitness identification evidence. *Journal of Applied Research in Memory and Cognition, 10,* 356-362. https://doi.org/10.1016/j.jarmac.2021.06.005

Smalarz, L., Kornell, N., Vaughn, K. E., & Palmer, M. A. (2019). Identification performance from multiple lineups: Should eyewitnesses who pick fillers be burned? *Journal of Applied Research in Memory and Cognition*, *8*, 221-232. https://doi.org/10.1016/j.jarmac.2019.03.001

Smalarz, L., & Wells, G. L. (2015). Contamination of eyewitness self-reports and the mistaken-identification problem. *Current Directions in Psychological Science*, *24*, 120-124. https://doi.org/10.1177/0963721414554394

Smith, A. M., Wilford, M., Quigley-McBride, A., & Wells, G. L. (2019). Mistaken identifications increase when either witnessing or testing conditions get worse. *Law and Human Behavior, 43,* 358-368. https://doi.org/10.1037/lhb0000334

Smith, A. M., Smalarz, L., Ditchfield, R., & Ayala, N. T. (2021). Evaluating the claim that high confidence implies high accuracy in eyewitness identification. *Psychology, Public Policy, and Law*, *27*(4), 479–491. https://doi.org/10.1037/law0000324

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41,* 127–145. https://doi.org/10.1037/lhb0000219

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. http://dx.doi.org/10.1037/0033-2909.118.3.315

Steblay, N. K., & Wells, G. L. (2020). Assessment of bias in police lineups. *Psychology, Public Policy, and Law, 26,* 393- 412. https://doi.org/10.1037/law0000287

Thomas Haynesworth v. Commonwealth of Virginia, Va. Ct. App. 0224112 (2011).

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

Valentine, T., Darling, S., & Memon, A. (2007). Do strict rules and moving images increase the reliability of sequential identification procedures? *Applied Cognitive Psychology, 21,* 933-949. https://doi.org/10.1002/acp.1306

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44,* 3-36. https://doi.org/10.1037/lhb0000359

Wells, G. L., & Olson, E. A. (2001). The other-race effect in eyewitness identification: What do we do about it? *Psychology, Public Policy, and Law, 7,* 230- 246. https://doi.org/10.1037/1076-8971.7.1.230

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

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*(5), 835–844. https://doi.org/10.1037/0021-9010.78.5.835

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: an experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior, 39,* 1-14. https://doi.org/10.1037/lhb0000096

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

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005

*Winsor, A. A., Flowe, H. D., Seale-Carlisle, T. M., Killeen, I. M., Hett, D., Jores, T., Ingham, M., Lee, B. P., Stevens, L. M., & Colloff, M. F. (2021). Child witness expressions of certainty are informative. *Journal of Experimental Psychology: General*, *150*(11), 2387–2407. https://doi.org/10.1037/xge0001049

Wixted, J. T. (2018). Time to exonerate eyewitness memory. *Forensic Science International, 292,* e13-e15. https://doi.org/10.1016/j.forsciint.2018.08.018

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309. https://doi.org/10.1073/pnas.1516814112

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science*, *13*, 324-335. https://doi.org/10.1177/1745691617734878

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10–65. https://doi.org/10.1177/1529100616686966

Table 1

Pristine Lineup Conditions (Reproduced from Box 3, Wixted & Wells, 2017)

1. Include only one suspect per lineup
2. The suspect should not stand out in the lineup
3. Caution that the offender might not be in the lineup
4. Use double-blind testing
5. Collect a confidence statement at the time of the identification

Table 2

ID Frequencies in Fair (A), Partially Biased (B), and Biased (C) Lineups

	Distribution of Lineup Misidentifications						Estimated Number of Innocent Suspect IDs		
Lineup	#1	#2	#3	#4	#5	#6	Nominal Size Correction	Effective Size Correction	No Correction
A	10	10	10	10	10	10	10	10	60
B	0	20	20	20	0	0	10	20	60
C	0	0	60	0	0	0	10	60	60

Note. Data are hypothetical. The nominal size correction divides total misidentifications by the number of lineup members. The effective size correction divides total misidentifications by the number of plausible lineup members. No correction represents the total number of misidentifications.

Table 3

Characteristics of the 10 Studies Published by Other Authors

Reference	Sample Size			Trials	IDs	Lineup Size	Filler Rotation	Filler Pool
	Culprit Present	Culprit Absent	Total					
Akan et al. (2021) Exp 1	1,321	1,296	2,617	1	2,617	4, 6, 8	Yes	91[2]
Colloff & Wixted (2020) Exp 3	555	555	555	2	1,110	6	Yes	34
Colloff et al. (2016)[3]	3,401	3,397	6,798	1	6,798	6	Yes	40
Colloff et al. (2021) Exp 1	5,263	5,296	10,559	1	10,559	6	Yes	110
Colloff et al. (2021) Exp 2	4,553	4,620	9,173	1	9,173	6	Yes	110
Nyman et al. (2019)	775	775	775	4	3,100	8	No	—
Seale-Carlisle et al. (2019) Exp 1	464	481	945	1	945	9	No	—
Seale-Carlisle et al. (2019) Exp 2	584	536	1,120	1	1,120	9	No	—
Seale-Carlisle et al. (2019) Exp 5	446	445	891	1	891	9	No	—
Winsor et al. (2021)	1,111	1,094	2,205	1	2,205	6	No	—

[2] The authors report using a pool of 64 fillers, but the unique filler numbers in the dataset total 91.
[3] The condition that was intentionally biased against the suspect was excluded.

Table 4

Effective Size of Culprit-Absent Lineups in Experiments 1—3

Exp	Culprit	Innocent Suspect	Fair	Biased
1	A	Best Match	4.43 [3.78, 5.35]	1.62 [1.35, 2.02]
2	A	Best Match	5.32 [4.58, 6.33]	1.71 [1.35, 2.34]
3	C	Weak Match	3.54 [3.13, 4.07]	1.59 [1.29, 2.09]
	D	Weak Match	3.71[1] [3.12, 4.55]	1.25 [1.09, 1.47]
		Next Best Match	3.71[1] [3.12, 4.55]	1.45 [1.20, 1.84]
		Best Match	3.71[1] [3.12, 4.55]	1.29 [1.12, 1.52]

[1]For Culprit D, the three innocent suspects for culprit-absent fair lineups were all in the same lineup.

Table 5

Lineup Identification Responses in Experiments 1—3

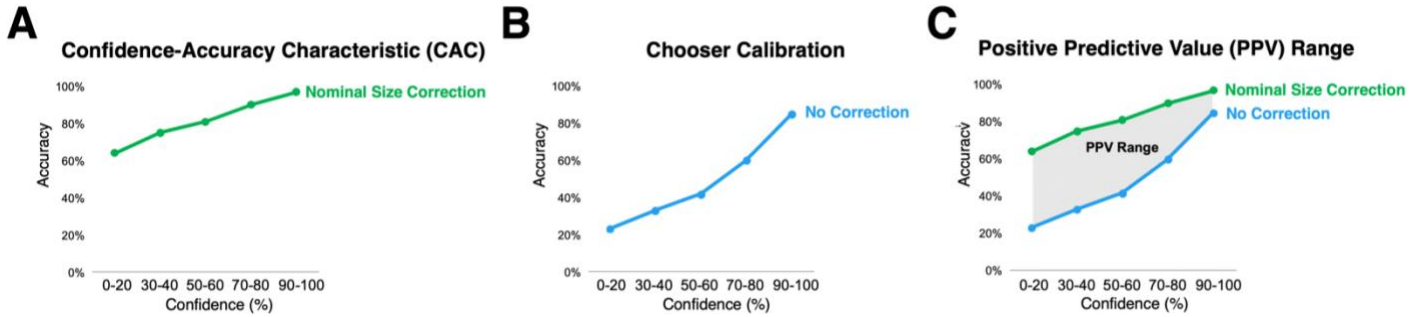
Lineup	Exp	Culprit	Innocent Suspect	Bias	Identification Response (%)			*N*
					Suspect	Filler	No ID	
Culprit Present	1	A	–	Biased	75.1	11.8	13.1	229
				Fair	59.2	28.0	12.8	218
	2	A	–	Biased	70.9	10.0	19.1	110
				Fair	40.7	34.5	24.8	113
	3	C	–	Biased	62.7	4.7	32.7	150
				Fair	19.0	43.5	37.4	147
		D	–	Biased	65.6	4.6	29.8	151
				Fair	22.4	42.9	34.6	156
Culprit Absent	1	B	Best Match	Biased	58.6	17.1	24.3	111
				Fair	25.0	48.4	26.6	128
	2	B	Best Match	Biased	45.7	14.9	39.4	94
				Fair	15.9	48.9	35.2	88
	3	C	Weak Match	Biased	29.7	8.2	62.0	158
				Fair	1.3	60.1	38.6	153
		D	Weak Match	Biased	42.6	5.2	52.3	155
				Fair	5.2	49.0	45.8	153[1]
			Next-Best Match	Biased	33.3	7.2	59.5	153
				Fair	20.3	34.0	45.8	153[1]
			Best Match	Biased	46.5	6.4	47.1	155

	Fair	17.0	37.3	45.8	153[1]

[1]For Culprit D, the three innocent suspects for culprit-absent fair lineups were all in the same lineup – rates for best match and description match innocent suspects were produced by varying the designated innocent suspect (i.e., from the same dataset) to match the innocent suspect in the biased lineups.

Figure 1

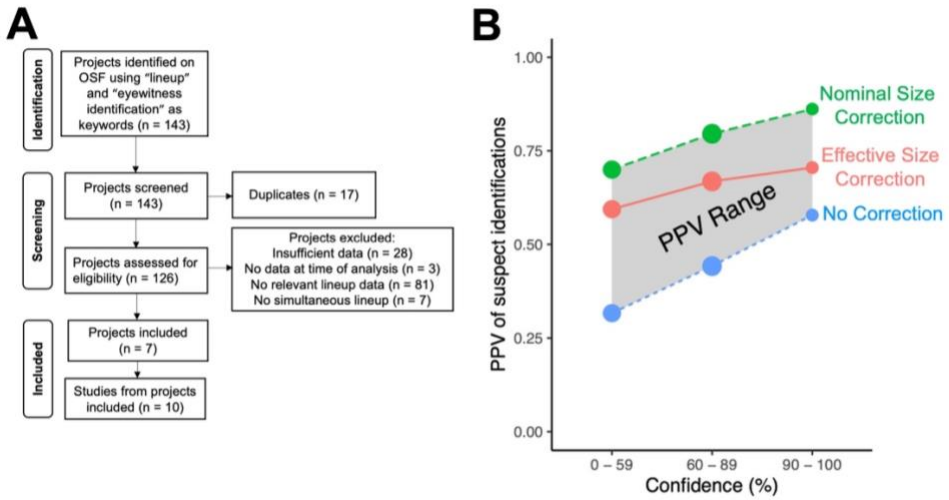
Accuracy and Confidence in Synthesis by Wixted and Wells (2017)



Note. Figure is adapted from Wixted and Wells (2017), who synthesized data from 15 studies. Each data point represents the accuracy of suspect identifications, or positive predictive value (PPV), under different assumptions. Panel A represents PPV if innocent suspect IDs are estimated by dividing the error rate in culprit-absent lineups by the lineup's nominal size, which assumes the lineups are perfectly fair. Panel B shows the suspect ID accuracy curve if no correction is applied to the false positive identification rate in culprit-absent lineups, which assumes every error from the culprit-absent lineup is an innocent suspect ID. In Panel C, which depicts both curves together, the shaded area represents the spectrum of possible PPVs across the lineup fairness continuum. We refer to this shaded area as the PPV range.

Figure 2

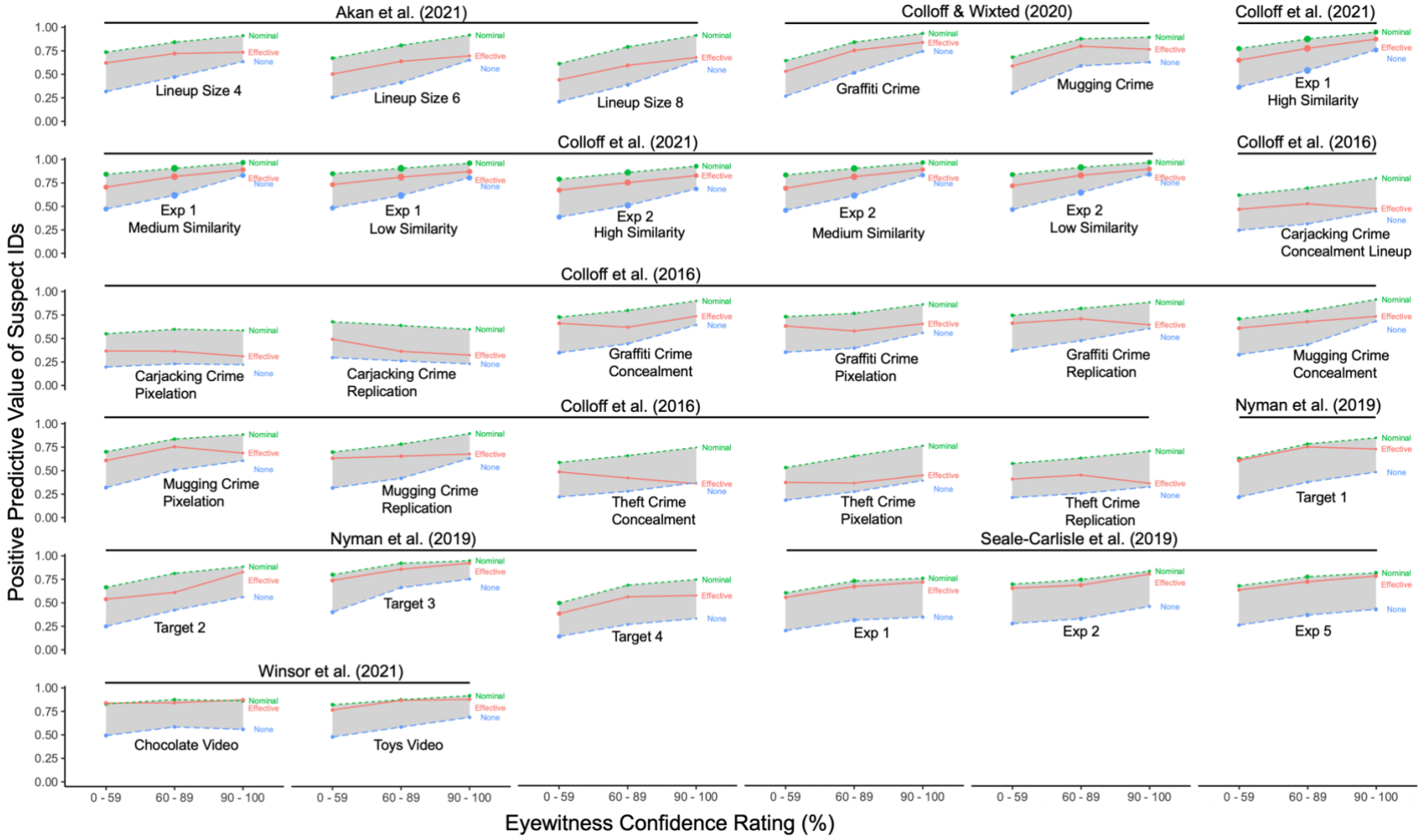
Meta-Analysis of 10 Previously Published Studies



Note. Panel A depicts the search for studies on the Open Science Framework (OSF) that report the distribution of mistaken IDs across culprit-absent lineup members (flow chart is adapted from the PRISMA template, Page et al., 2021). Panel B depicts summary estimates of the Positive Predictive Value (PPV) of suspect IDs, depending on the eyewitness confidence rating and the method of estimating innocent suspect IDs (nominal size correction, effective size correction, or no correction). Circles are proportional to sample sizes.

Figure 3

PPV Ranges in the 10 Previously Published Studies



Note. The nominal size correction assumes the lineup is perfectly fair to the suspect. The effective size correction factors in lineup fairness, based on the distribution of mistaken IDs in the culprit-absent lineup. No correction assumes the lineup is maximally biased and that every false positive in the culprit-absent lineup is a mistaken identification of an innocent suspect. The shaded area represents the PPV range across the lineup fairness continuum. Circles are proportional to sample sizes.

Figure 4

Culprits, Designated Innocent Suspects, and Lineups in Experiments 1—3

Note. Stimulus sets in Experiments 1–3. The lineup images were recorded using a booth on loan from the Video Identification Parade Electronic Recording (VIPER) Bureau, West Yorkshire Police, England. These images have not been quality assured by the VIPER Bureau, and the authors accept full responsibility for their quality. The people depicted are actors, not actual culprits or lineup members in real criminal cases. All actors consented to publication of their photograph in academic journal articles.

Figure 5

Effect of Lineup Bias on Correct IDs in Experiments 1—3

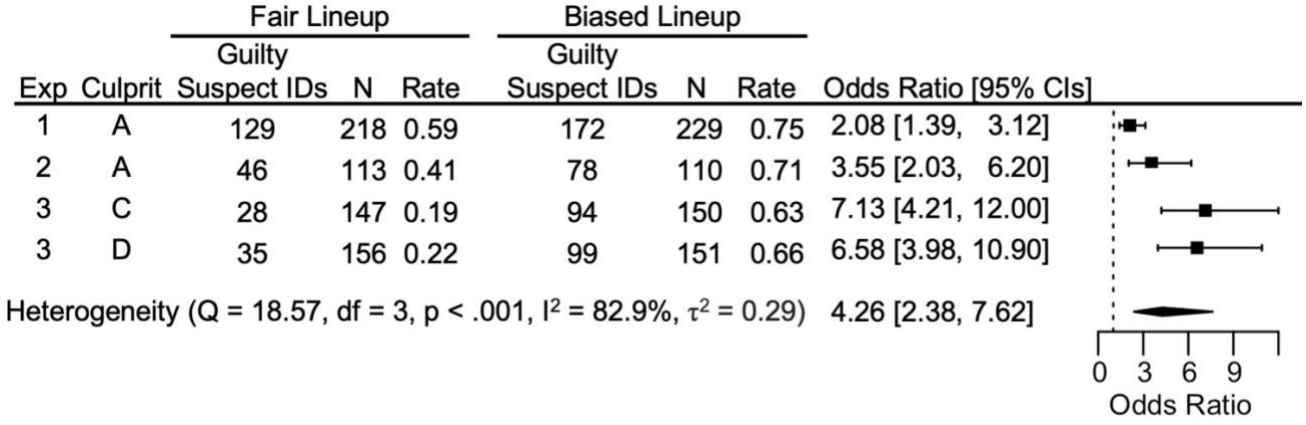


Figure 6

Effect of Lineup Bias on Innocent Suspect IDs

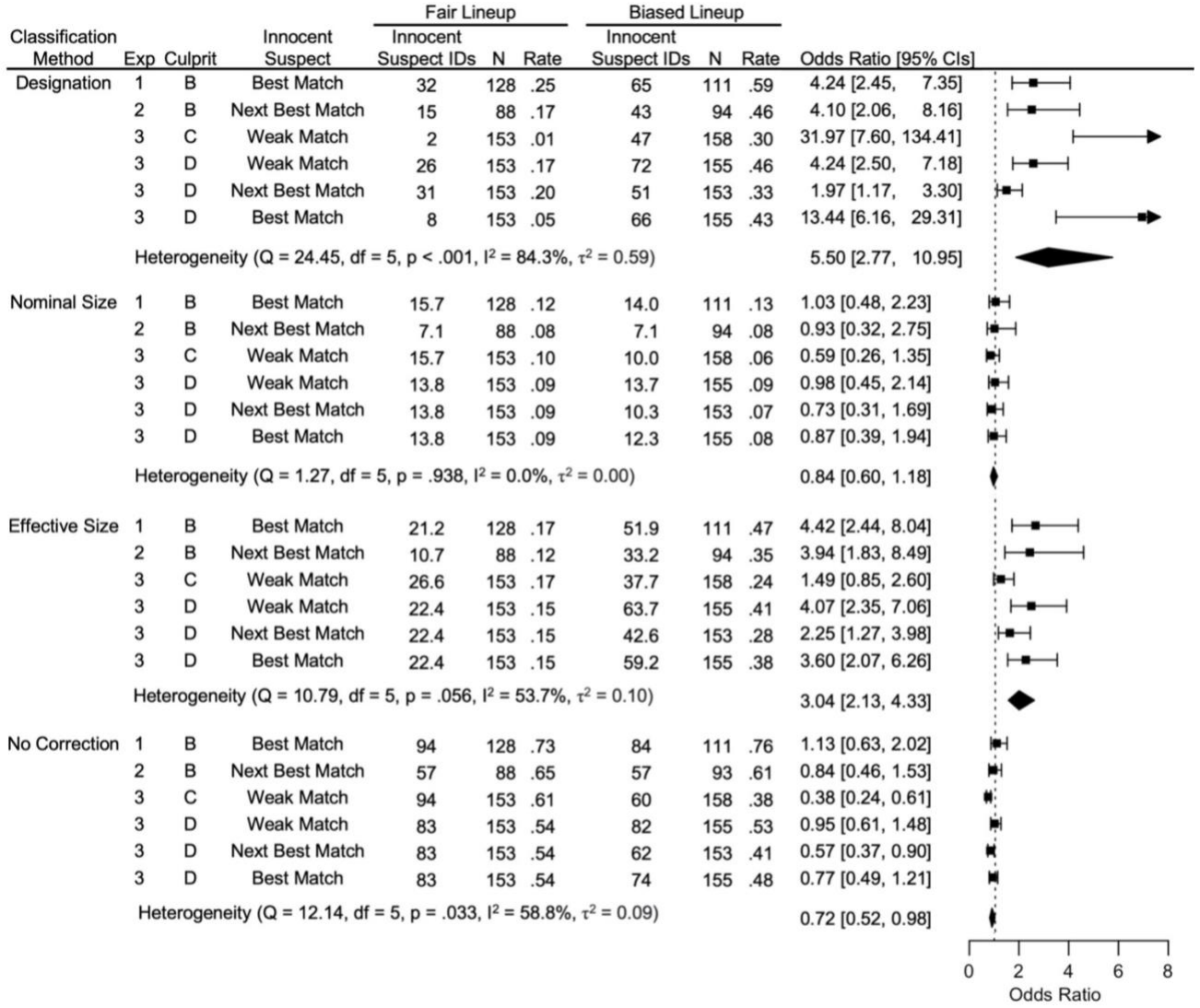
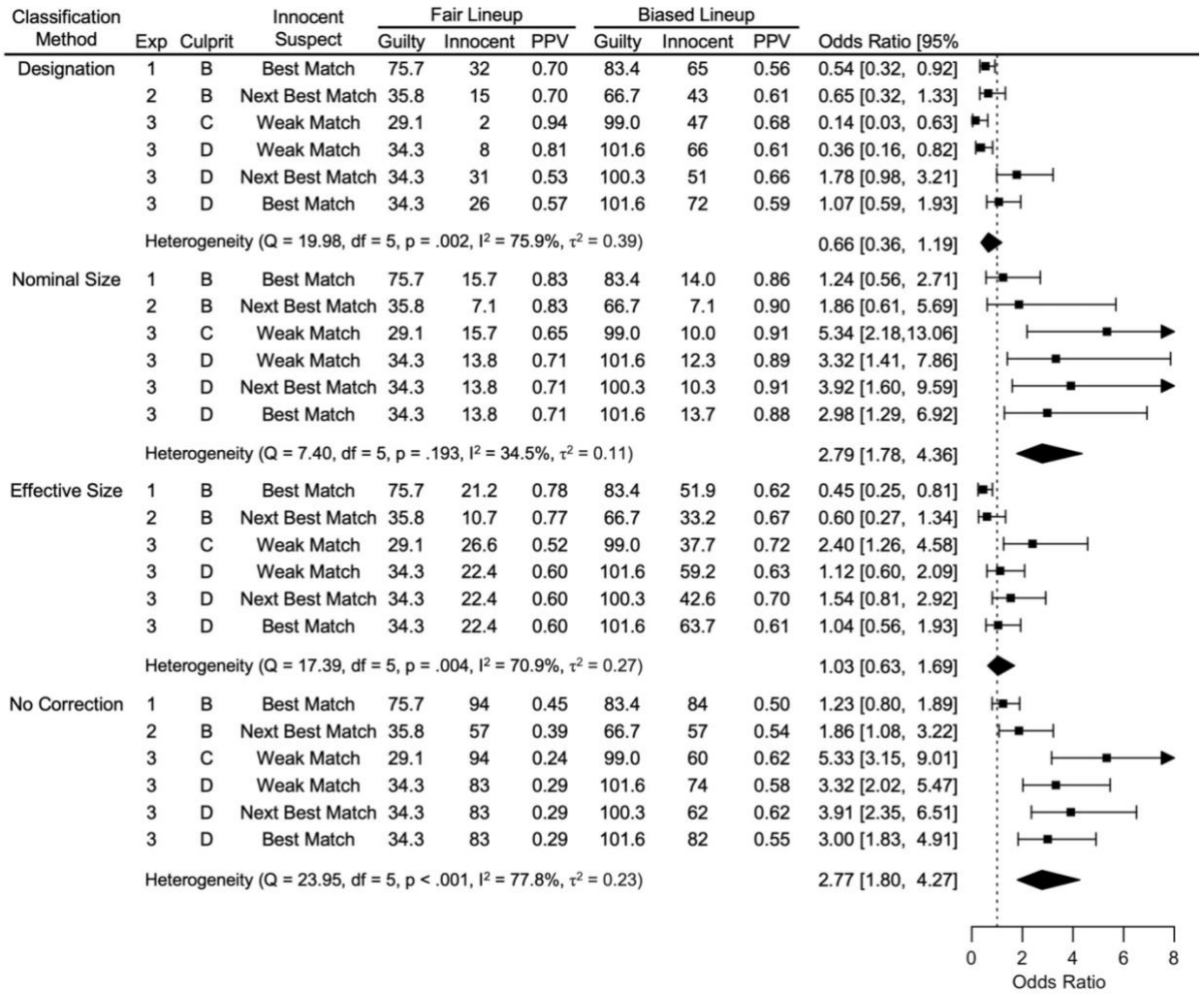


Figure 7

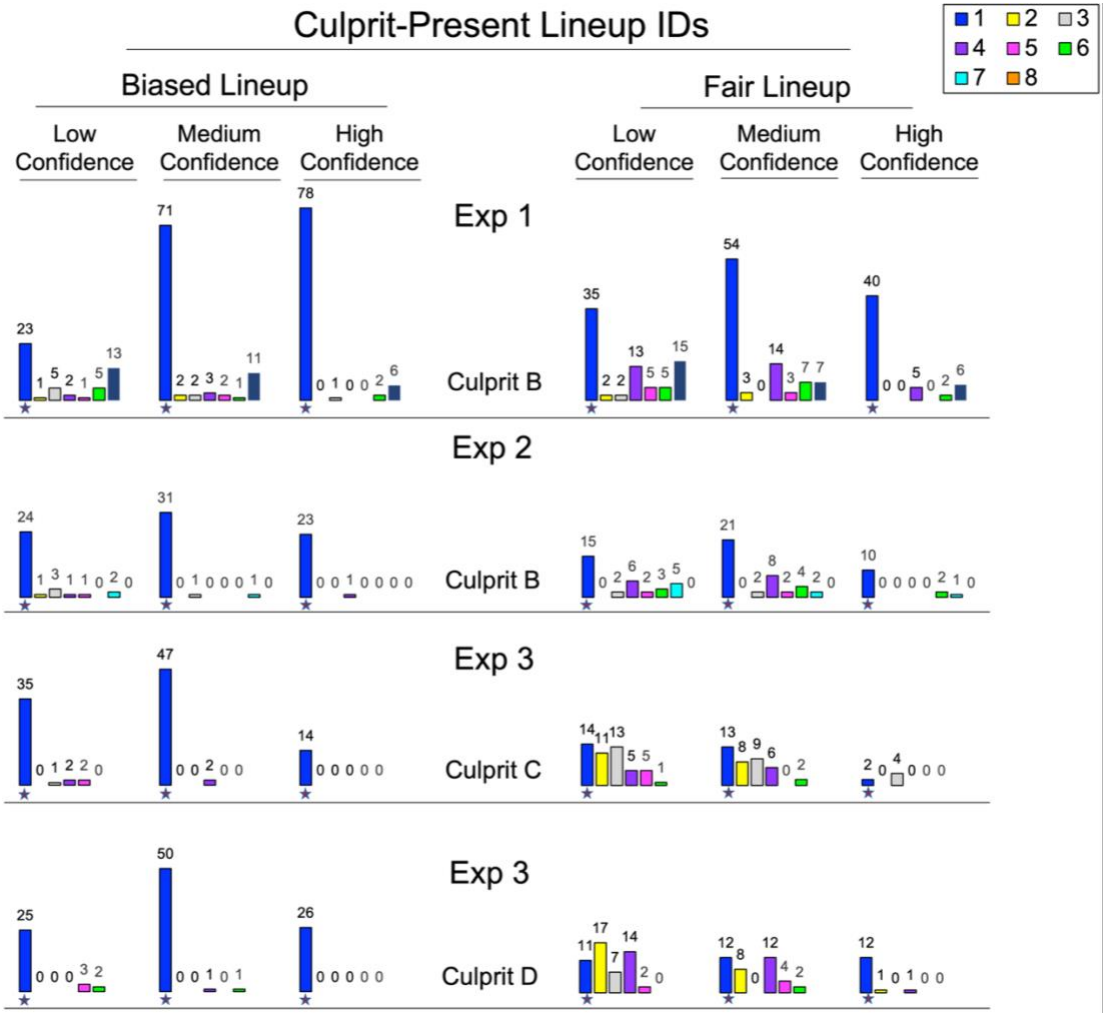
Effect of Lineup Bias on the Positive Predictive Value (PPV) of Suspect IDs



Note. Values for Guilty and Innocent are suspect ID frequencies. Guilty suspect IDs are corrected to equate culprit-present and culprit-absent lineup sample sizes.

Figure 8

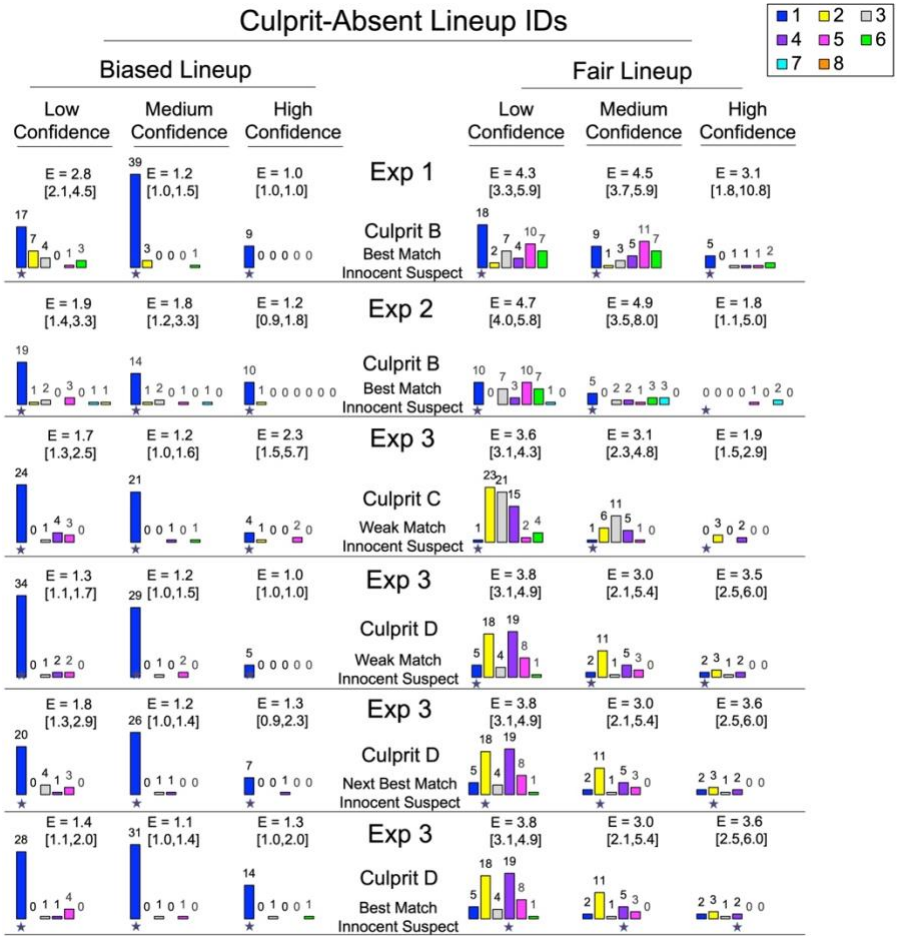
Distribution of identifications across the culprit-present lineups in Experiments 1—3



Note. Low confidence = 0-59%, moderate confidence = 60-89%, and high confidence = 90-100%. Bars represent individual lineup members, with ID numbers corresponding with those assigned in Figure 1. Culprits are indicated with a star (★).

Figure 9

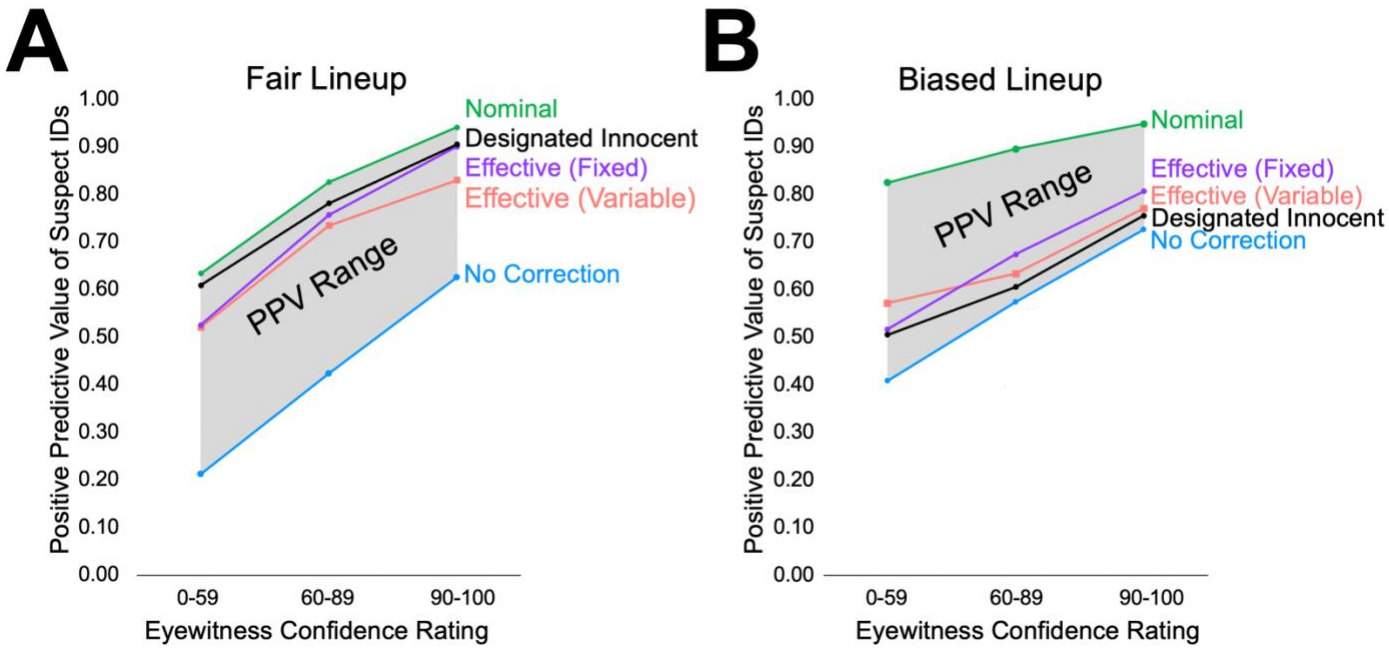
Distribution of Identifications Across the Culprit-Absent Lineups in Experiments 1—3



Note. Low confidence = 0-59%, moderate confidence = 60-89%, and high confidence = 90-100%. E = Effective Size, with scores that could range from $E = 1$ (maximally biased) to $E = k$ (perfectly fair), where k is the number of lineup members (Exp 1 = 6, Exp 2 = 8, Exp 3 = 6). Values in square brackets are 95% CIs. Bars represent individual lineup members, with ID numbers corresponding with those assigned in Figure 4. Designated innocent suspects are indicated with a star. The three plots for Culprit D depict the same data, with different designated innocent suspects.

Figure 10

Positive Predictive Value (PPV) Range in Fair and Biased Lineups in Experiments 1—3



Note. The nominal size correction represents performance, assuming the lineup is perfectly fair. No correction represents performance, assuming all culprit-absent lineup members are innocent suspects. The effective size corrections represents performance, assuming the suspect is among the plausible lineup members.