Estimation of Eyewitness Error Rates in Fair and Biased Lineups

[Supplemental Materials]

Table of Contents

Effective Size Calculations

For the re-analysis of previous studies, two authors independently extracted the datasets from the OSF and independently calculated an estimate of effective size (*E*) for all subgroups (using Microsoft Excel, or the R package r4lineups [Tredoux & Naylor, 2018]) . For the studies that used the same lineups within all subgroups (Nyman et al., 2019; Seale-Carlisle et al., 2019; Winsor et al., 2021), calculation of *E* was straightforward and computed as described by Tredoux (1998), except data from witnesses in the lineup experiment were used in place of data from non-witnesses in a Mock Witness Procedure (Quigley-McBride & Wells, 2021). Although stimulus sampling in the remaining studies was an asset for avoiding stimulus-specific effects, it also complicated calculation of *E*. Specifically, because fillers for 6-member lineups were randomly selected from larger pools of fillers, each lineup member's identification probability had to be weighted to account for the number of times that the filler appeared in a lineup.

For studies that used the same lineups within all subgroups, *E* was calculated from the distribution of identification responses using the procedure described by Tredoux (1998): In Step 1 the number of identifications for each lineup member was extracted directly from the study's OSF datafile. In Step 2, an identification proportion was computed separately for each lineup member:

$$P_i = \frac{O_i}{N}$$

where P_i is the identification proportion for a given lineup member (i.e., the identification rate), O_i is the observed identification frequency for the lineup member, and N is the total number of identifications of all lineup members.

In Step 3, the index of diversity (*I*) was computed to give a measure of how much the observed distribution of choices deviates from a uniform choice distribution (Agresti & Agresti, 1978):

$$I = 1 - \sum_{i=1}^{k_l} (P_i)^2$$

where P_i and N are as defined above and k_l is the number of lineup members.

Finally, in Step 4, the index of diversity (I) was transformed into Tredoux's (1998) measure of lineup effective size, applied to the observed choice distributions:

$$E = \frac{1}{1 - I}$$

The E calculation is demonstrated with an example from the Winsor et al. (2021) dataset. The distribution of choices and associated probabilities for the low confidence subgroup that saw a culprit-absent lineup labeled "chocolate" are presented in Table S1. From this distribution, $I = 1 - .187 = .813$, and, $E = 1/(1 - .813) = 5.347$. The interpretation of an effective size of 5.35 is that identifications were distributed across more than five of six of the lineup members.

Table S1

Example data for estimating effective size of culprit absent lineup (E)

Lineup Member	ID Frequency	ID Proportion (P_i)	P_i^2
1	14	.144	.021
2	19	.196	.038
3	12	.124	.015
4	7	.072	.005
5	22	.227	.051
6	23	.237	.056
Sum =	97	1.000	.187

Weighting in Filler Rotation Studies

In filler rotation studies, fillers for 6-member lineups are randomly selected from larger pools of fillers. To compute an effective size estimate in filler rotation studies, we weighted each lineup member's identification proportion to account for the number of times that filler was selected to appear in a lineup. To illustrate, Table S2 displays hypothetical data for three fillers selected from a larger pool

of fillers. Although all three fillers were identified 10 times, Filler 1 appeared in lineups less often than did Filler 2. Therefore, the ratio of identifications to appearances was larger for Filler 1, and it is reasonable to infer that Filler 1 was a more attractive choice than Filler 2. By the same logic, Filler 2 was more plausible than Filler 3. To correct the observed identification frequencies for the number of appearances, we first computed a relative identification proportion by dividing the observed identification frequency by the number of lineup appearances (e.g., Filler 1 = 10/50 = .20). Next, we needed to generate weights that would adjust the observed identification frequency for the number of appearances without affecting the total identification frequency. These weights were computed by multiplying each filler's relative identification proportion by the sum of all lineup members' relative identification proportions (e.g., Filler 1 = .20/.51 =.39). Finally, a corrected identification frequency was computed by multiplying each lineup member's weight by the sum of the observed identification frequencies (e.g., Filler 1 = .39 × 30 = 11.77). This procedure accounted for the number of lineup appearances while keeping the sums of the observed and corrected identification frequencies constant.

Table S2

Hypothetical data for weighting the number of identifications by the number of lineup member appearances in studies that sampled lineup fillers from filler pools.

Filler	Lineup Appearance Frequency	Observed Identification Frequency	Relative Identification Proportion	Weight	Corrected Identification Frequency
1	50	10	.20	.39	11.77
2	60	10	.17	.33	9.82
3	70	10	.14	.28	8.41
Sum	180	30	.51	1.00	30.00

The corrected identification frequencies were used to compute the effective size of the filler pool, E_p. Following the four steps described above, we computed E from the observed choice distributions. The effective size of the filler pool was then transformed into the effective size of the lineup, E, with the formula:

$$E = \frac{E_p}{k_p}(k_l)$$

where k_p is the number of fillers in the pool and k_l is the number of fillers in the lineup, which we standardized to five for target-present lineups and six for culprit-absent lineups (assuming a 6-person lineup).

This produced a value that was comparable to the estimate of E when fillers were not sampled from larger pools.

Variance estimates were derived in order to compute confidence intervals, which were computed using the approach recommended by Tredoux (1998) (when computing them in Excel, but when computing them in R we estimated 95% bootstrap intervals instead). Namely, the variance of I, v_I, was computed with the formula:

$$v_I = \frac{4}{N}\left\{\sum_{i=1}^{k_l}(P_i)^3 - \left[\sum_{i=1}^{k_l}(P_i)^2\right]^2\right\}$$

where P_i is the identification proportion for a given lineup member (i.e., the rate of identification), N is the total number of identifications of all lineup members, and k_l is the number of lineup members.

Then 95% confidence intervals for E were generated (where 1.96 is the $Z_{.975}$ value for the normal distribution):

$$LL = \frac{1}{\left\{1 - \left(I - 1.96({v_I}^{0.5})\right)\right\}}$$

$$UL = \frac{1}{\left\{1 - \left(I + 1.96({v_I}^{0.5})\right)\right\}}$$

where I and E are as described above.

Finally, standard error and variance (for the meta-analysis) were generated

$$SE = \frac{LL - UL}{3.92}$$

and

$$v = SE^2$$

Experiment 1 Programming Error

After collecting data from the preregistered target sample size in Experiment 1, a programming error was discovered. Specifically, the image of the banana was mistakenly omitted from the culprit-absent video. Thus, the study was terminated for participants in the culprit-absent condition unless they correctly guessed the answer to the manipulation check. This resulted in a disproportionate number of participants in the culprit-present condition. When the issue was discovered, the experiment was relaunched to collect the target number of participants in the culprit-absent condition. Thus, assignment to the culprit-present and culprit-absent conditions was not entirely balanced. However, within the culprit-present and culprit-absent conditions, participants were randomly assigned to the fair or biased lineup conditions (and to the similarity conditions).

Participant Exclusions

In Experiment 1, data from 1120 participants were excluded because they did not complete the experiment (n = 287) or they did not meet the inclusion criteria (n = 833). Most of the incomplete cases (n = 285) made minimal progress through the experiment and we had no identification data from them to analyze. The remaining two participants completed the experiment but did not respond to the categorical identification decision question. The following predefined inclusion criteria were used in Experiment 1: 18+ years old, fluent in English, normal or corrected-to-normal vision, and using a computer or laptop. Participation was terminated immediately, and respondents were exited from the experiment, if any of the following occurred: they did not to consent to participate (n = 106), they reported they were not 18+

years (n = 43), they reported problems with the video (n = 56), they reported their vision was not normal or corrected to normal (n = 104), they reported they were not fluent in English (n = 40), they reported they were using a mobile device (n = 22), they failed the 6-item multiple choice attention check question (n = 462).

In Experiment 2, the Qualtrics recruitment service supplied us with a datafile that excluded respondents who did not meet the following inclusion criteria: Consented to participate, participated from the United States, participated on a PC or Laptop, 18+ years old, fluent in English, Had normal or corrected to normal vision. The Qualtrics service also excluded participants who failed the 6-item multiple choice attention check question or reported technical problems with the video.
After these exclusions, the datafile included 1211 participants. For the present research, we excluded data from participants who received 4- member (n = 404) or 6-member lineups (n = 402) and only analyzed data from participants who received the 8-member lineups (n = 405).

In Experiment 3, we received responses from 1647 participants. We excluded 86 participants whose data were incomplete and 30 participants who failed the 6-item multiple choice attention check question, resulting in a final sample of 1531.

Participant Demographics

The final sample of Experiment 1 included 363 men, 302 women, 1 participant who responded "other", and 1 who responded "prefer not to say." All participants were required to be 18 years or older, but specific age was not collected. Table S3 shows most participants self-identified as White (86%), and only a small proportion self-identified as Black (6%). We checked to see if the Black participants happened to be disproportionately assigned to any condition, given that the actors in the videos and lineups were also Black and this could have given the Black participants a same-race advantage on the identification task. In Experiment 1, 19 Black participants were assigned to the fair lineup and 19 were

assigned to the biased lineup, though the distribution was less equal when culprit-presence was included as a factor (fair: culprit present = 15, culprit absent = 4; biased: culprit present = 12, culprit absent = 7).

The final sample of Experiment 2 included 197 men, 204 women, 2 "other", and 2 "prefer not to say." All participants were required to be 18 years or older. Specific age was not collected. Table S3 shows that most participants self-identified as White (76%), and only a small proportion self-identified as Black (12%). We checked to see if the Black participants happened to be disproportionately assigned to any condition, which showed that 27 Black participants were assigned to the fair lineup (culprit present = 14, culprit absent = 13) and 21 were assigned to the biased lineup (culprit present = 12, culprit absent = 9).

The final sample of Experiment 3 included 731 men, 778 women, 5 participants who responded "other", and 17 who responded "prefer not to say." Participants ages ranged from 18-84 years (M = 42.0, SD = 13.0). Self-reported ethnicity is reported in Table S3.

Table S3

Self-Identified Ethnicity or Race of Participants

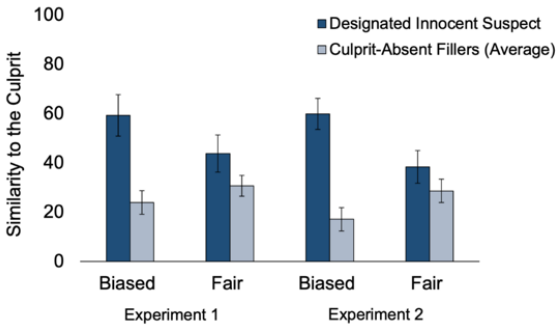
	Exp 1		Exp 2		Exp 3	
Ethnicity	*n*	%	*n*	%	*n*	%
White	570	85.5	306	75.5	1139	74.4
Black	38	5.7	48	11.9	145	9.5
Asian	30	4.5	21	5.2	114[1]	7.5
Hispanic/Latin American	1	0.1	18	4.4	84	5.5
Mixed	14	2.1	7	1.7	17	1.1
Other	2	0.3	3	0.7	8	0.5
Prefer not to say	12	1.8	2	0.5	24	1.6

[1] In Exp 3 we collected more fine-grained information for the Asia category, which included Chinese (*n* = 39), Southeast Asian (*n* = 30), South Asian (*n* = 24), Korean (*n* = 17), and Japanese (*n* = 4).

Similarity Ratings

Figure S1

Similarity of Culprit-Absent Lineup Members to the Appearance of the Culprit



Note. Similarity was rated on a scale from 0% to 100%. Error bars are 95% confidence intervals.

In Experiment 1, half of the participants in Experiment 1 were randomly assigned to rate each lineup member's similarity to their memory of the culprit on a scale from 0% (not similar) to 100% (similar) *before* making a categorical lineup identification decision. Figure S1 shows that average ratings for the designated innocent suspect were higher than average ratings for the fillers and that the difference between the innocent suspect and the fillers was greater in biased lineups than in fair lineups. Similarity ratings for each lineup member are reported in Table S4.

In Experiment 2, half of the participants in Experiment 2 were randomly assigned to rate each lineup member's similarity to their memory of the culprit on a scale from 0% (not similar) to 100% (similar) *after* making a categorical lineup identification decision. These ratings are reported in Table S4. Figure S1 shows that average ratings for the designated innocent suspect were higher than average ratings for the fillers and that the difference between the innocent suspect and the fillers was greater in biased lineups than in fair lineups.

Table S4

Rated Similarity of Lineup Members to Memory of the Culprit

		Exp 1						Exp 2					
		Culprit Present			Culprit Absent			Culprit Present			Culprit Absent		
Lineup	Member	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
Biased	1 (Suspect)	76.6	29.6	115	59.4	31.5	54	70.9	30.3	110	60.0	31.6	94
	2 (Filler)	14.4	20.6	115	25.4	26.5	54	16.3	26.0	110	18.7	26.0	94
	3 (Filler)	20.6	26.3	115	25.0	26.3	54	21.9	29.1	110	20.5	27.0	94
	4 (Filler)	16.3	22.4	115	23.5	23.8	54	21.5	27.6	110	17.7	26.1	94
	5 (Filler)	14.1	21.2	115	19.4	22.1	54	16.5	27.1	110	13.8	24.3	94
	6 (Filler)	17.6	22.0	115	26.5	26.4	54	23.5	29.6	110	20.4	28.0	94
	7 (Filler)	-	-	-	-	-	-	17.7	27.9	110	15.4	26.8	94
	8 (Filler)	-	-	-	-	-	-	18.9	27.6	110	13.6	23.4	94
	Filler Avg	16.6	18.6	115	24.0	18.0	54	19.5	25.3	110	17.2	23.5	94
Fair	1 (Suspect)	74.0	28.7	109	43.9	30.5	62	57.1	32.4	113	38.4	32.0	88
	2 (Filler)	22.5	23.8	109	21.8	22.9	62	20.0	25.9	113	20.7	26.4	88
	3 (Filler)	16.6	21.2	109	28.7	29.1	62	19.5	25.9	113	34.2	30.7	88
	4 (Filler)	42.5	33.0	109	31.6	25.8	62	34.8	32.4	113	32.0	30.2	88
	5 (Filler)	23.1	24.6	109	38.4	29.5	62	21.9	24.7	113	32.8	30.5	88
	6 (Filler)	30.8	28.5	109	33.4	29.1	62	22.7	25.0	113	26.7	27.7	88
	7 (Filler)	-	-	-	-	-	-	25.0	27.5	113	25.7	31.6	88
	8 (Filler)	-	-	-	-	-	-	25.4	29.9	113	28.9	28.5	88
	Filler Avg	27.1	18.9	109	30.8	16.9	62	24.2	21.3	113	28.7	22.4	88

Note. The numbers assigned to lineup members in this table correspond with the numbers assigned to lineup members in Figure 4 in the main article.

Pilot Study 1

Before launching Experiment 1, a pilot test was conducted with a small group of participants recruited from the Qualtrics survey panel (total $n = 73$, after incompletes and terminations $n = 49$). After analysing the data from these participants, some minor adjustments were made to the stimuli. For the pilot test we used the same crime video for the culprit-present condition as in the main experiment but we used a different crime video for the culprit-absent condition (i.e., not Culprit B in Figure 4 of the main article). These participants identified the guilty suspect (62%) only slightly more often than they identified the innocent suspect (52%). Thus, to minimize the risk of a floor effect (i.e., inability to discriminate between guilty and innocent suspects), we used someone less similar to the culprit for the culprit-absent crime video (i.e., Culprit B, Figure 4) and to improve the quality of the fair lineup we replaced one lineup member with the person who appeared in the culprit-absent crime video in the pilot study (#6 in the fair lineup, Figure 4). The pilot study data were not included in the analyses reported in the main paper.

Pilot Study 2

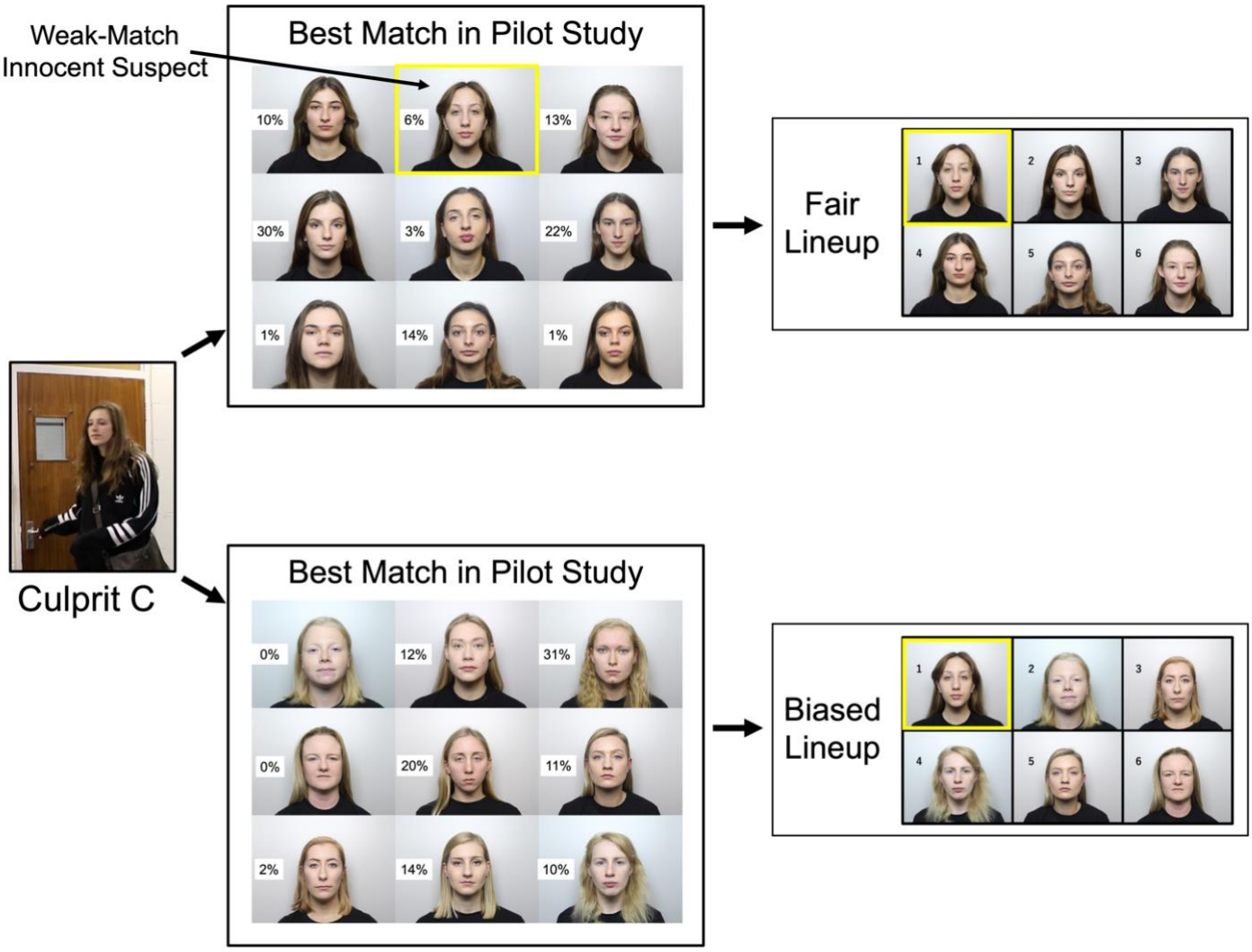
Prior to collecting data for Experiment 3, we conducted pilot research to measure the match between fillers and innocent suspects to the culprits. The General Procedure for the experiments was followed in the pilot research: Participants viewed a crime video with Culprit C or Culprit D, completed an attention check, played the Snakes game for 3 mins, provided a description of the culprit, and completed a lineup task. Culprit-absent lineups depicted nine women with light brown hair or nine women with blonde hair. For the lineup task, undergraduate students ($N = 520$) selected the person who best matched their memory of the culprit, indicated if the person they selected was the culprit, and rated their confidence. Culprit-absent data ($n = 386$) are reported in Figures S2 and S3. Culprit-present data ($n = 134$) were not used for lineup construction and are not discussed further.

Our goal in constructing the lineups was to maximize the strength of the lineup fairness manipulation. Using responses to the best match question for culprit-absent lineups, we eliminated the three least frequently selected light-brown-haired lineup members, who were consistent for participants who saw Culprit C and Culprit D (see Figures S2 and S3). This left us with a suspect and five fillers who all had light brown hair for the fair lineup.

From the 9-member blonde lineup, we needed to eliminate four lineup members so that we would have five fillers for the biased lineup. We initially planned to eliminate the four most frequently selected blonde-haired lineup members (to maximize bias toward the innocent suspect). Had we followed this initial plan, one filler would have differed across the lineups for Culprit C and Culprit D (see Figures S2 and S3). This led us to weigh the benefit toward our goal of maximizing bias from having a potentially less competitive lineup member in the biased lineup in relation to the benefit toward the goal of experimental control from using the same fillers across culprits. We decided the benefit of a potentially less competitive lineup member would likely be negligible, given that any description-mismatched filler would be unlikely to compete strongly with an innocent suspect who matched the culprit's description. Therefore, we opted for experimental control and included the 3rd most frequently identified lineup member from participants who saw Culprit D. Apart from this deviation, we followed our initial plan to eliminate the most frequently selected blonde-haired lineup members (see Figures S2 and S3).

Figure S2

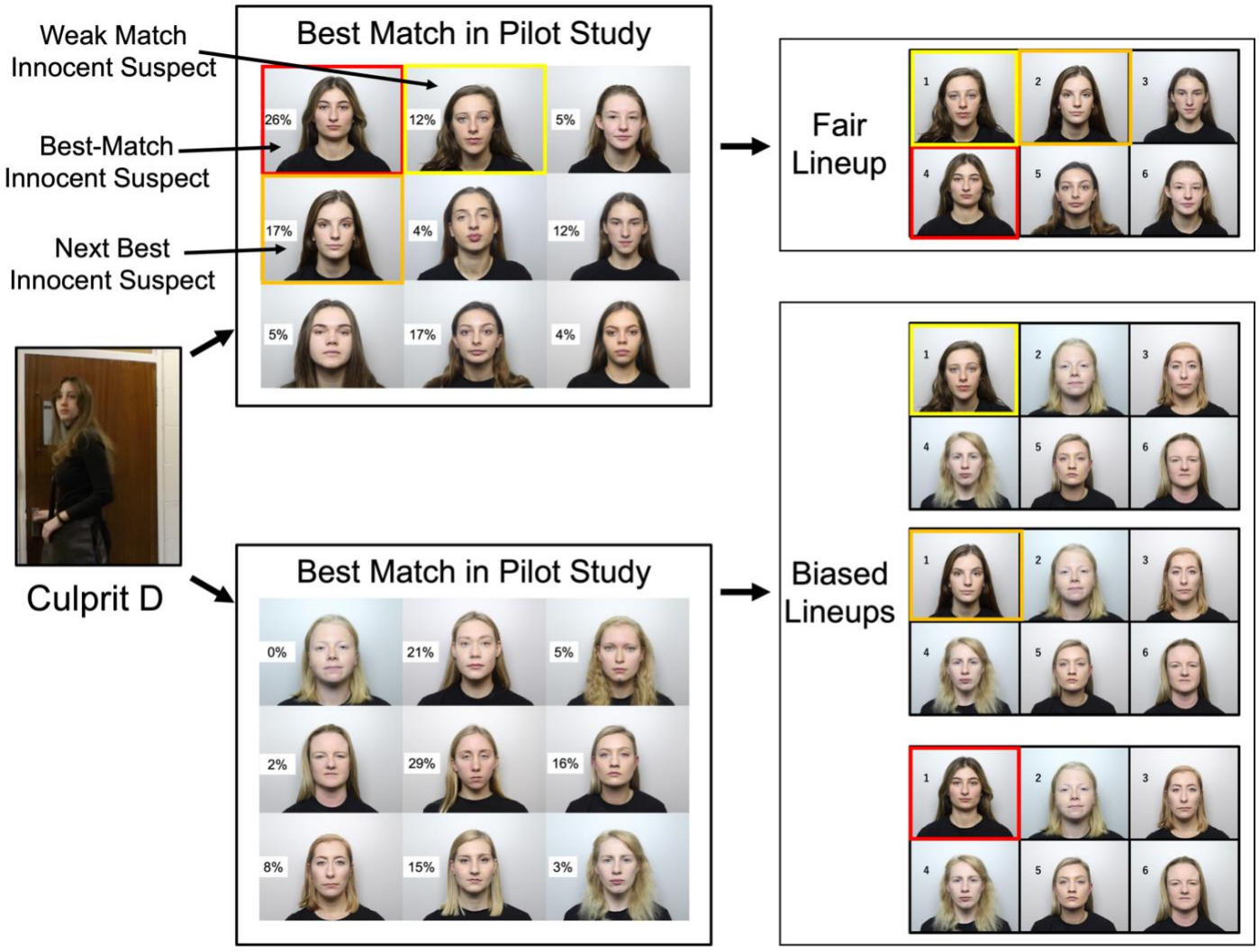
Filler selection process for culprit-absent lineups for Culprit C in Experiment 3



Note. 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. Sample size is $n = 84$ for the blonde lineup and $n = 87$ for the light-brown-hair lineup.

Figure S3

Filler selection process for culprit-absent lineups for Culprit D in Experiment 3



Note. 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. Sample size is n = 131 for the blonde lineup and n = 84 for the light-brown-hair lineup.

References

Agresti, A., & Agresti, B. F. (1978). Statistical analysis of qualitative variation. *Sociological methodology*, *9*, 204-237.

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

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.* Taylor and Francis.

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

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

Tredoux, C.G., & Naylor, R. (2018). r4lineups: Statistical inference on lineup fairness. https://cran.r-project.org/web/packages/r4lineups/index.html

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