Supplemental Materials:

Live presentation for eyewitness identification is not superior to photo or video presentation

Methodological Limitations in Previous Research

We reviewed previous research that involved a live identification test and coded the method sections for (1) the number of targets that were used for the witnessed event; (2) an explicit statement indicating that participants were randomly assigned to the live and nonlive conditions; and (3) sample size in each condition. These data are reported in Table SM1.

Experiment 1

Method

Participants. In Experiment 1a, we recruited 318 participants. Data from 14 participants were excluded due to live presentation error (five participants), presence at multiple exposure events (four participants), previous encounter with the target (two participants), and age below 15 years (three participants). The final sample consisted of 304 participants, 89 men and 211 women (two participants indicated other and two participants did not disclose their sex). Participants self-reported their ethnicity as White (n = 266), Asian (n = 23), Black (n = 6), Mixed (n = 1), other (n = 3), and prefer not to say (n = 1).

In Experiment 1b, the total sample included 310 participants. Data from four participants were excluded due to technical fault at administration (n = 1), live presentation error (the suspect wore glasses; n = 1), and previous encounter with the target (n = 2). The final sample consisted of 306 participants, 78 men and 226 women (one participant indicated other and one did not disclose their sex). Participants self-reported their ethnicity as White (n = 240), Asian (n = 18), Black (n = 26), Mixed (n = 14), other (n = 5), and prefer not to say (n = 3).

Table SM1

Number of targets, explicit reporting of random assignment (RA), and sample sizes in previous studies that included live identification tests

				Target Present			Target Absent			
No.	Study	Target(s)	RA	Live	Video	Photo	Live	Video	Photo	*N*
1	Cutler & Fisher 1990	1 man, 1 woman	yes	26	26	21	26	26	21	146
2	Cutler, Fisher, & Chicvara 1989	1 man	no	17	16	NA	9	8	NA	50
3	Dent & Stephenson 1979 (Exp 2)	1 man	no	98	NA	124	NA	NA	NA	222
4a	Dent & Stephenson 1979 (Exp 4 screen)	1 man	no	50	NA	50	NA	NA	NA	100
4b	Dent & Stephenson 1979 (Exp 4 no screen)	-	-	50	NA	-	NA	NA	NA	50
5	Egan, Pittner, & Goldstein 1977	2 men	no	40	NA	46	NA	NA	NA	86
6	Kerstholt, Koster, van Amelsvoort 2004	1 man	no	58	48	44	58	49	45	302
7a	Peters 1991 (crime)	1 man	yes	12	NA	12	12	NA	12	48
7b	Peters 1991 (no crime)	1 man		12	NA	12	12	NA	12	48
8	Shepherd, Ellis, & Davies 1982 (Exp 1)	4 persons	no	12	80	NA	NA	NA	NA	92
9	Shepherd, et al. 1982 (Exp 4 live vs video)	2 men	yes	19	20	29	NA	NA	NA	68
10	Sporer 1991 (sim)	1 person	yes	15	NA	13	13	NA	15	56
11	Sporer 1991 (seq)	1 person	yes	11	NA	13	12	NA	15	51
12	Valentine, Davis, Memon, & Roberts 2012 (Exp 1)	1 woman	no	47	95	NA	48	93	NA	283
13	Valentine et al. 2012 (Exp 2)	1 man	no	49	114	NA	NA	NA	NA	163
14	Valentine et al. 2012 (Exp 3)	1 woman	no	125	93	NA	92	96	NA	406
	Total/Average	1.31	36%	641	492	364	282	272	120	2171

Note. NA = Not Applicable

Anxiety. Following a confidence rating for the identification decision in Experiment 1a, we asked: “How anxious were you while making the identification decision?” with a 7-point Likert scale (‘1 = not at all anxious’, and ‘7 = very anxious’).

Motivation manipulation. At the first few data collection events of Experiment 1a, we offered incentives to suspects to deter them from being identified. On a portion of the trials, suspects were informed that they would earn a bonus if they were identified less than expected. These data and analyses are not reported here because this manipulation was abandoned part way through data collection due to our conclusion that it was ineffective.

Analysis

The target sampling employed in all experiments reported in this paper created a specific challenge when deciding on data analyses. There was an obvious option of collapsing data across targets, or an alternative of taking target-level effects into account and employing multi-level and meta-analytical approaches. In addition, there are multiple ways of analyzing eyewitness identification data (e.g., Lampinen, 2016; Mickes, Moreland, Clark, & Wixted, 2014; Wells, Smalarz, & Smith, 2015). We opted to employ a variety of different analytic approaches to determine whether they would yield a consistent conclusion (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

In the following, we first present multi-level analyses of hit rates, false alarm rates, accuracy, and choosing (measures conceptually similar to those reported in the main article) using a series of generalized linear mixed models (GLMMs) built with the `lme4` package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2016). These models included the conditions of interest in the fixed part of the model. Due to the use of multiple targets across events, we nested participants' responses into targets in the random part of the model. Medium effects were analyzed in two models using simple contrasts with the medium factor, so that the first model compared live vs photo and live vs video conditions, and the second model compared photo vs video conditions.

For analyses of confidence, we explored how confidence and the medium related to accuracy in choosers and non-choosers separately (Wells & Penrod, 2011). However, the patterns of results were consistent irrespective of choosing; therefore, we report results for the whole sample.

All significant interactions were followed-up with models for separate groups. For all binomial tests, we report odds ratios (OR) along with 95% confidence intervals (CI) in square brackets to show the range of plausible values for the size of the effect (Cumming, 2012, 2014). In cases where the OR is below 1, we report the inverse of the values and add a corresponding verbal description to aid interpretation of the effect size. We computed Cohen's *d* as an approximate (collapsed across targets) measure of effect size for continuous dependent variables.

Hits. Target identifications were coded as hits and all other target-present decisions were treated as non-hits (i.e., misses and don't know responses were collapsed).

False alarms. Identifications of the innocent suspect were coded as false alarms and all other target-absent decisions were not treated as false-alarms (i.e., correct rejections and don't know responses were collapsed).

Accuracy. In target-present conditions, suspect identifications were treated as accurate responses and all other decisions were treated as inaccurate responses (i.e., misses and don't know responses were collapsed). In target-absent conditions, rejections were treated as accurate responses and all other decisions were treated as inaccurate responses (i.e., false alarms and don't know responses were collapsed).

Choosing. Participants who made an identification decision were treated as choosers; participants who made any other decision were treated as non-choosers (rejections and don't know responses were collapsed).

Results: Experiment 1a

Hits. There were significantly more hits in the non-live showups than in the live showups (photo vs live: $OR = 2.34$ [1.09, 4.99], $z = 2.19$, $p = .028$; video vs live: $OR = 2.52$ [1.14, 5.57], $z = 2.29$, $p = .022$). There were no significant differences in hit rates between the photo and video showups ($OR = 1.08$ [0.47, 2.47], $z = 0.18$, $p = .856$).

False alarms. There were no significant differences in false alarms between the live and photo showups (*OR* = 3.10 [0.87, 11.03], *z* = 1.75, *p* = .081). However, the live and video showup contrast indicated the video showup led to significantly more false alarms (*OR* = 3.56 [1.07, 11.79], *z* = 2.08, *p* = .038). False alarm rates did not differ significantly between the photo and video showups (*OR* = 1.14 [0.41, 3.20], *z* = 0.25, *p* = .802).

Accuracy. Analyses of accuracy with medium as a predictor indicated no significant effects (highest *z* = 1.07, lowest *p* = .287).

Choosing. Irrespective of accuracy, participants made more identifications from photo and video showups than from live showups (live vs photo: *OR* = 2.65 [1.39, 5.06], *z* = 2.95, *p* = .003; live vs video: *OR* = 2.76 [1.45, 5.27], *z* = 3.08, *p* = .002). There were no significant differences in choosing between the photo and video showups (*OR* = 1.07 [0.56, 2.06], *z* = 0.21, *p* = .833).

Confidence, medium, and accuracy. The analysis involving live vs non-live medium comparisons indicated that higher confidence was associated with higher accuracy (*OR* = 1.33 [1.10, 1.60], *z* = 2.97, *p* = .003). There were no significant effects of medium (highest *z* = 1.16, lowest *p* = .245) and no interactions (highest *z* = 1.20, lowest *p* = .232).

The model involving only the photo and video medium comparison indicated a similar relationship between confidence and accuracy (*OR* = 1.27 [1.001, 1.60], *z* = 2.02, *p* = .044). The effect of the medium and the interaction between confidence and medium were not significant (highest *z* = 1.23, lowest *p* = .218).

Anxiety and medium. The first model that included the live vs non-live comparisons revealed a main effect of medium in the live vs video contrast and two interactions between target presence and the live vs photo and live vs video contrasts. We followed-up these interactions by splitting the sample into target-present and target-absent conditions.

In target-present conditions, anxiety ratings were lower in the video (*M* = 2.26, *SD* = 1.21) than in the live condition (*M* = 3.39, *SD* = 1.72; *d* = 0.74 [0.32, 1.15]; *b* = -1.13, *SE* = 0.30, *t*(149) =

3.78, $p < .001$). Anxiety ratings were also lower in the photo ($M = 2.88$, $SD = 1.37$) than in the live condition, but the difference was not significant ($d = 0.32$ [-0.06, 0.71]; $b = -0.51$, $SE = 0.28$, $t(149) = 1.80$, $p = .075$). There were no significant differences in anxiety ratings between target-absent live and non-live conditions (highest $t = 1.34$, lowest $p = .18$).

The second model including the photo vs video contrast for medium and target presence revealed a significant main effect of medium. Participants in the video condition reported lower anxiety ratings ($M = 2.54$, $SD = 1.44$) than participants in the photo condition ($M = 2.94$, $SD = 1.41$; $d = 0.28$ [-0.02, 0.57]; $b = -0.41$, $SE = 0.21$, $t(181) = 1.98$, $p = 0.049$).

Anxiety, medium, choosing, and accuracy. In this exploratory analysis, we looked at how medium, choosing, and accuracy related to anxiety. The first model that included the live vs non-live comparisons revealed a significant three-way interaction between live vs video contrast for medium, choosing, and accuracy. We followed up this interaction by separate analyses for choosers and non-choosers in the live and video conditions.

In the live condition, low numbers of choosers precluded any statistical comparisons (there were only five choosers who made an inaccurate decision). In non-choosers, there was a significant difference in anxiety between participants who made an inaccurate compared to an accurate decision ($b = -1.02$, $SE = 0.34$, $t(76) = 2.98$, $p = .004$). Participants who made incorrect rejections reported higher levels of anxiety ($M = 3.43$, $SD = 1.69$) than participants who made correct rejections ($M = 2.41$, $SD = 1.32$; $d = 0.67$ [0.21, 1.14]).

In the video condition, choosers who made a correct identification reported lower levels of anxiety ($M = 2.35$, $SD = 1.29$) than choosers who made a false identification ($M = 3.38$, $SD = 1.66$; $d = 0.73$ [0.02, 1.44]; $b = -1.04$, $SE = 0.48$, $t(37) = 2.15$, $p = .038$). The pattern of the anxiety-accuracy relationship in non-choosers in the video condition was opposite to the one found in the live condition: participants who made a correct rejection reported higher levels of anxiety ($M = 2.85$, $SD = 1.59$) than

participants who made an incorrect rejection (*M* = 1.96, *SD* = 1.00; *d* = 0.67 [0.09, 1.24]; *b* = 0.89, *SE* = 0.38, *t*(49)= 2.37, *p* = .022).

The second full model that included the photo vs video medium comparison revealed a significant two-way interaction between choosing and accuracy. There were no significant differences in reported anxiety for choosers who made accurate (*M* = 2.55, *SD* = 1.33) and inaccurate decisions (*M* = 3.13, *SD* = 1.60) across both non-live showups (*d* = 0.41 [-0.08, 0.91]; *b* = -0.58, *SE* = 0.35, *t*(79)= 1.67, *p* = .100). Across both non-live showups, non-choosers who made an accurate rejection reported higher levels of anxiety (*M* = 3.03, *SD* = 1.50) than non-choosers who made an inaccurate rejection (*M* = 2.42, *SD* = 1.32; *d* = 0.43 [0.03, 0.83]; *b* = 0.61, *SE* = 0.28, *t*(102)= 2.17, *p* = .032).

Results: Experiment 1b

Hits. We found no significant differences in hit rates across the three pairs of showup comparisons (live vs photo: *OR* = 1.29 [0.60, 2.80], *z* = 0.65, *p* = .518; live vs video: *OR* = 1.88 [0.80, 4.39], *z* = 1.45, *p* = .147; photo vs video: *OR* = 1.55 [0.65, 3.68], *z* = 0.99, *p* = .325).

False alarms. Similarly, we found no significant differences in false alarm rates (live vs photo: *OR* = 2.20 [0.60, 8.05], *z* = 1.19, *p* = .234; live vs video: *OR* = 1.39 [0.36, 5.39], *z* = 0.48, *p* = .631; photo vs video: *OR* = 0.62 [0.19, 2.03], *z* = 0.79, *p* = .432).

Accuracy. Analyses of accuracy with medium as a predictor indicated no significant effects (highest *z* = 1.80, lowest *p* = .072).

Choosing. There were no significant differences in choosing across the three pairs of medium comparisons (highest *z* = 1.35, lowest *p* = .177).

Confidence, medium, and accuracy. The model comparing live vs photo and live vs video showups showed that participants who were more confident tended to also be more accurate (*OR* = 1.67, 95% *CI* [1.35, 2.05], *z* = 4.83, *p* < .001). There were no significant effects of medium (highest *z* = 1.20, lowest *p* = .231) and no interactions (highest *z* = 1.52, lowest *p* = .128).

The model comparing photo vs video showups revealed a significant interaction: the relationship between confidence and accuracy was stronger in the video condition (video: *OR* = 2.32 [1.49, 3.62], $z = 3.70$, $p < .001$) than in the photo condition (*OR* = 1.32 [0.98, 1.80], $z = 1.80$, $p = .072$).

Results: Experiment 1a and 1b

Confidence-accuracy characteristics for Experiment 1a and 1b. The accuracy of suspect identifications as a function of confidence was plotted as a Confidence Accuracy Characteristic (CAC) curve (Mickes, 2015). Figure SM1 displays levels of confidence on the x-axis and percentage of correct suspect identifications on the y-axis. Each point represents the proportion of suspect identifications that were accurate at a given level of confidence; the dashed grey reference line represents perfect calibration. Although we interpret these data with caution (few decisions were made at the highest level of confidence; see Table SM2), Figure SM1 indicates relatively poor calibration and little differences across the media.

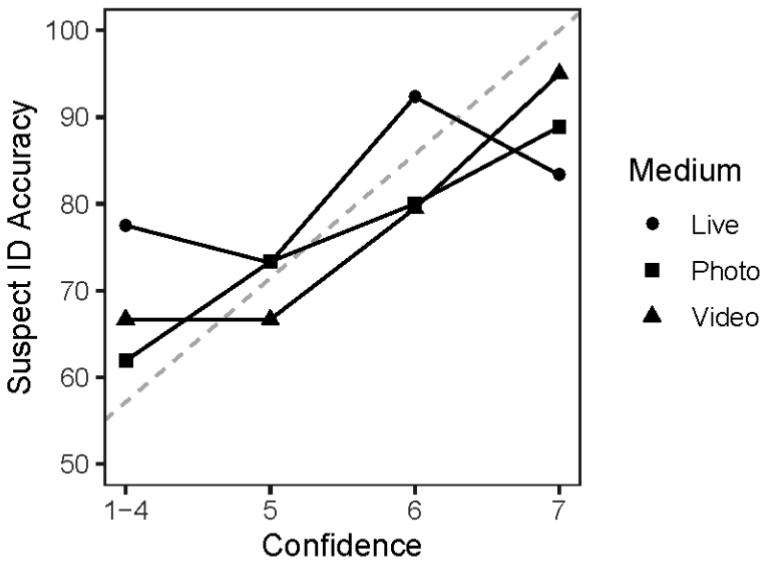

Figure SM1. Confidence-accuracy characteristic curve for data from Experiments 1a and 1b. Lower levels of confidence (1-4) were collapsed due to low frequency.

Table SM2

Frequency of identification decisions at different levels of confidence

	Live		Photo		Video	
Confidence	Hit	False Alarm	Hit	False Alarm	Hit	False Alarm
1 – 4	7	2	8	4	10	6
5	11	4	17	8	14	5
6	25	2	17	5	16	6
7	3	1	15	1	11	2

Receiver Operating Characteristic (ROC) analyses. Another way of comparing eyewitness identification procedures is via computing a series of diagnosticity ratios based on confidence ratings, plotting these values, and then comparing areas under the curves to see which procedure yields higher diagnosticity (see Gronlund, Wixted, & Mickes, 2014). Diagnosticity ratios for each medium were computed by systematically removing identifications made at lowest levels of confidence in a stepwise fashion. Figure SM2 displays these values plotted for the live, photo, and video conditions: the rightmost points represent diagnosticity ratios including identifications made at all levels of confidence (i.e., 1 – 7) the next point to the left represents the values computed when identifications made with the lowest level of confidence were removed (i.e., 2 – 7) etc., and the leftmost point represents values computed for identifications made with the highest confidence (i.e., 7). Partial areas under the ROC curves (pAUC) were then computed and compared using the `pROC` package in *R* (Robin et al., 2011). No significant differences were detected (live vs photo: $D = 0.28$, $p = .782$; live vs video: $D = 0.51$, $p = .611$; photo vs video: $D = 0.26$, $p = .796$).

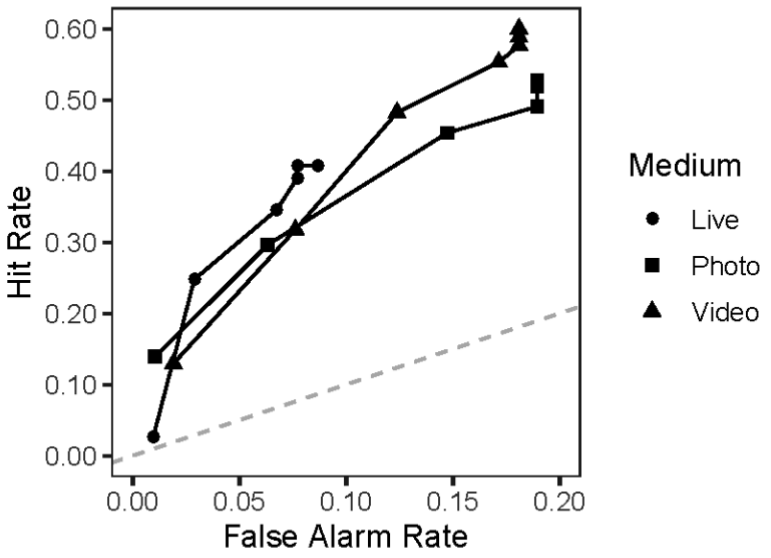

Figure SM2. Receiver operating characteristic curves for live, photo, and video conditions in Experiments 1a and 1b. The grey line represents chance performance.

Discussion

In Experiment 1a, analyses employing nested models indicated one significant effect that was inconsistent with analyses of data collapsed across targets: there was a higher false-alarm rate in the video than in the live condition. All other analyses were consistent with results reported in the main article.

Experiment 2

Method

Participants. Out of the 438 participants, 436 were presented with both the male and the female lineup; two participants in the live condition viewed only the female lineup. We excluded partial data from 18 participants due to programming errors (n = 11 in the video conditions saw either the target or the innocent suspect as the first lineup member), live presentation error (n = 4), technical errors (n = 2), or incomplete data (n = 1). The final sample consisted of 856 lineup decisions from

438 participants, 112 men and 326 women. Participants in the final sample self-reported their ethnicity as White (*n* = 360), Black (*n* = 27), Mixed (*n* = 24), Asian (*n* = 20), and prefer not to say (*n* = 7).

Statistical analyses. In multi-level analyses, we added a random factor into all GLMMs that nested responses into participants to reflect that each participant made two lineup decisions.

Results

Overview. After presenting additional data relevant for resultant lineup fairness, we report supplementary analyses of hits and false alarms. Then, we present CAC curves, ROC analyses, and Maximum Utility and Deviation from Perfect Performance (DPP). Finally, we describe exploratory analyses of: (i) comfort ratings, (ii) pre- and post-identification confidence ratings, (iii) endorsements of features used to make an identification decision, (iv) decision justifications, (v) repeated lineups viewings, and (vi) not sure responses.

Resultant lineup fairness. Table SM3 shows the distribution of identifications across fillers and resultant lineup fairness (Tredoux's *E'*) in target-absent lineups for each stimulus set computed using the `r4lineups` package (Tredoux & Naylor, 2018).

Table SM3

Proportions of filler choices in target-absent lineups

Lineup Gender	Stimulus Set	Person						*n*	*E'*
		1	2	3	4	5	6*		
Male	a	.29	.00	.14	.14	.29	.14	7	4.45
	b	.22	.00	.11	.11	.00	.56	9	2.61
	c	.11	.00	.17	.00	.56	.17	18	2.66
	d	.31	.00	.15	.08	.39	.08	13	3.60
	e	.60	.20	.00	.20	.00	.00	5	2.27
	f	.17	.00	.17	.08	.25	.33	12	4.24
	g+	.00	.00	.00	.00	.00	.00	0	NA
	h	.20	.00	.00	.00	.40	.40	5	2.78
Female	a	.00	.11	.00	.22	.33	.33	9	3.52
	b	.14	.21	.14	.07	.21	.21	14	5.44
	c	.29	.14	.21	.00	.14	.21	14	4.67
	d	.20	.26	.13	.20	.13	.07	15	5.23
	e	.00	.19	.06	.19	.31	.25	16	4.27
	f	.20	.10	.20	.10	.10	.30	10	5.00
	g	1.00	.00	.00	.00	.00	.00	1	1.00
	h	.00	.00	.00	.00	1.00	.00	3	1.00

Note. * = the designated innocent suspect. + = there were no identifications made for male stimulus set g (there were only rejections; $n = 3$).

Hits. In all GLMMs, suspect identifications were treated as hits and any other decisions were treated as non-hits (i.e., rejections and filler identifications were collapsed). The models revealed no significant differences in hit rates (live vs mugshot video: $OR = 1.10$ [0.67, 1.81], $z = 0.39$, $p = .700$; live vs full-body video: $OR = 1.30$ [0.79, 2.15], $z = 1.02$, $p = .306$; mugshot vs full-body video: $OR = 1.40$ [0.43, 1.16], $z = 1.37$, $p = .171$).

False alarms. We first present an analysis that complements results reported in the main text (i.e., analyses collapsed across targets). Here, we separated innocent suspect identifications, filler identifications, and rejections in target-absent lineups. We found no significant differences in decision

rates across the conditions, $\chi_2(4, N = 421) = 3.27$, $p = .514$, and evidence for the null hypothesis was very strong, $BF_{01} = 630.21$.

Next, we report two sets of GLMMs that also revealed consistent results. When only identifications of the designated innocent suspect were treated as false alarms, we found no significant differences across the conditions (live vs full-body video: *OR* = 2.37, [0.16, 1.13], $z = 1.72$, $p = .085$; live vs mugshot video: *OR* = 1.55 [0.23, 1.77], $z = 0.85$, $p = .394$; mugshot vs full-body video: *OR* = 1.99, [0.69, 3.37], $z = 1.04$, $p = .301$). Similarly, when all filler identifications were treated as false alarms, we found no significant differences across the conditions (live vs full-body video: *OR* = 1.14, [0.52, 1.46], $z = 0.52$, $p = .606$; live vs mugshot video: *OR* = 1.15, [0.53, 1.43], $z = 0.55$, $p = .580$; mugshot vs full-body video: *OR* = 1.01, [0.62, 1.59], $z = 0.02$, $p = .983$).

Confidence-accuracy characteristics. Figure SM3 displays CAC curves for the three conditions and Table SM4 shows frequency of identification decisions at different levels of confidence. As in Experiment 1, Figure SM3 indicates relatively poor calibration and little differences across the conditions.

Table SM4

Frequency of identification decisions at different levels of confidence

	Live		Mugshot Video		Full-Body Video	
Confidence	Hit	False Alarm	Hit	False Alarm	Hit	False Alarm
0 – 40%	6	4	7	7	5	9
50 – 60%	16	16	12	25	13	17
70 – 80%	11	17	15	18	18	16
90 – 100%	9	3	13	8	14	9

Note. Confidence-accuracy characteristic curves for Experiment 2 were calculated using the estimated innocent suspect identification rate (all mistaken identifications [innocent suspect + fillers]/6).

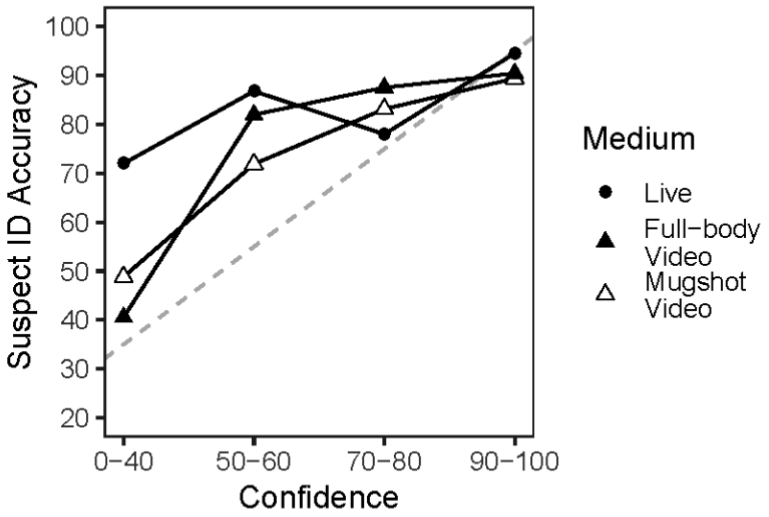

Figure SM3. Confidence-accuracy characteristic curve for Experiment 2. False alarms were computed as 1/6 of all filler identifications from target-absent lineups. Lower levels of confidence were collapsed due to low frequency.

ROC analyses. Diagnosticity analyses using pAUC (Figure SM4) revealed no significant differences among the conditions when only the designated innocent suspect identifications were treated as false alarms (live vs full-body video: $D = 0.26$, $p = .793$; live vs mugshot video: $D = 0.08$, $p = .939$; full-body vs mugshot video: $D = 0.03$, $p = .977$; see Panel A). When filler and innocent suspect identifications were treated as false alarms, full-body video had significantly higher diagnosticity than the live condition, $D = 2.03$, $p = .042$; other comparisons did not reveal significant differences (full-body vs mugshot video: $D = 1.46$, $p = .144$; live vs mugshot video: $D = 0.77$, $p = .440$; see Panel B).

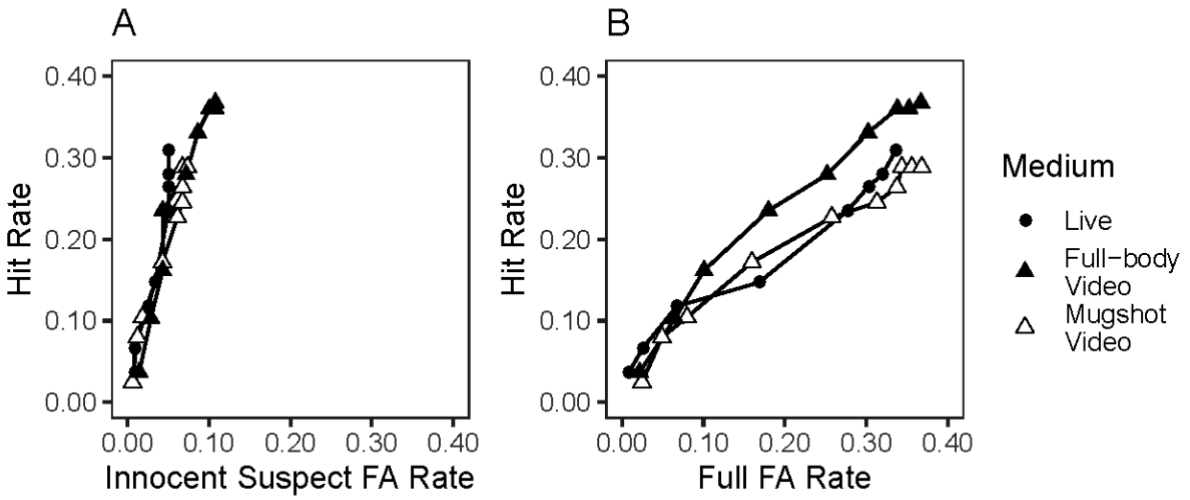

Figure SM4. Receiver-operating characteristic curve for the live, full-body video and mugshot video in Experiment 2 showing innocent suspect false alarm rate (A) and full false alarm rate (B). The grey line represents chance performance. FA = false alarm.

Maximum Utility and Deviation from Perfect Performance. Each operating point that comprised the ROC curve B in Figure SM4 were assessed for Maximum Utility (Clark, 2012; Lampinen et al., 2018) and DPP (Smith, Lampinen, Wells, Smalarz, & Mackovichova, 2019). Table SM5 shows that on average, the full-body video lineup showed the smallest deviation from perfect performance (0.81) than the live lineup (0.83) and the mugshot video lineup (0.85). However, none of the pairwise inferential comparisons were significant (95% confidence intervals of the difference in DPP values for full-body video vs live: -0.03 [-0.11, 0.05], full-body vs mugshot video: -0.04 [-0.13, 0.05], live vs mugshot video: 0.01 [-0.03, 0.05]).

Table SM5

Hits, false alarms, Deviation from Perfect Performance and Maximum Utility measures

	Live				Full-Body Video				Mugshot Video			
Confidence	Hits	FA	DPP	U	Hits	FA	DPP	U	Hits	FA	DPP	U
100%	.04	.01	0.97	0.03	.04	.01	0.98	0.02	.02	.01	0.98	0.02
90%	.07	.01	0.94	0.06	.10	.03	0.93	0.07	.08	.01	0.94	0.07
80%	.12	.03	0.91	0.09	.16	.04	0.88	0.12	.10	.02	0.91	0.09
70%	.15	.03	0.89	0.11	.24	.04	0.81	0.19	.17	.04	0.87	0.13
60%	.24	.05	0.82	0.18	.28	.07	0.79	0.21	.23	.06	0.83	0.17
50%	.26	.05	0.79	0.21	.33	.09	0.76	0.24	.25	.07	0.82	0.18
40%	.28	.05	0.77	0.23	.36	.10	0.74	0.26	.26	.07	0.80	0.20
30%	.31	.05	0.74	0.26	.36	.11	0.75	0.25	.29	.07	0.78	0.22
20%	.31	.05	0.74	0.26	.37	.11	0.74	0.26	.29	.07	0.78	0.22
10%	.31	.05	0.74	0.26	.37	.11	0.74	0.26	.29	.07	0.79	0.21
Average			0.83	0.17			0.81	0.19			0.85	0.15

Note. FA = false alarm (i.e., the designated suspect identification). DPP = Deviation from Perfect Performance. U = utility.

Accuracy. In target-present conditions, we treated hits as accurate decisions; filler identifications and misses were treated as inaccurate decisions. In target-absent conditions, we treated rejections as accurate decisions; filler identifications were treated as inaccurate decisions. Analyses of accuracy with medium as a predictor indicated no significant effects (highest $z = 1.02$, lowest $p = .308$).

Choosing. Participants who made an identification decision (hit or filler identification) were treated as choosers; participants who made a rejection decision were treated as non-choosers. We found no significant differences in choosing across the conditions (live vs full-body video: $OR = 1.07$ [0.66, 1.32], $z = 0.41$, $p = .683$; live vs mugshot video: $OR = 1.00$ [0.72, 1.40], $z = 0.02$, $p = .982$; mugshot vs full-body video: $OR = 1.08$ [0.78, 1.49], $z = 0.46$, $p = .646$).

Confidence, medium, and accuracy. The analyses indicated a negligible association between confidence and accuracy (*OR* = 1.19, [1.06, 1.34], *z* = 3.00, *p* = .003). There were no significant effects of medium or interactions (highest *z* = 0.61, lowest *p* = .544).

Comfort. In the post-lineup questionnaire, we asked participants: "How comfortable were you while making the identification decision?" with a percentage scale labeled at extreme values ('0% = very uncomfortable', and '100% = very comfortable'). Ratings of comfort were significantly lower in target-absent conditions (*M* = 72.02, *SD* = 24.46) than in target-present conditions (*M* = 75.18, *SD* = 23.25), but the effect was small (*d* = 0.13 [-0.002, 0.27], *b* = -2.84, *SE* = 0.93, *t*(435.42) = 3.03, *p* = .003). As for the medium, participants reported lower comfort ratings in both video conditions (mugshot: *M* = 72.25, *SD* = 23.83; full-body: *M* = 71.37, *SD* = 23.91) than in the live condition (*M* = 77.80, *SD* = 23.49; live vs full-body video: *d* = 0.27 [0.10, 0.44], *b* = -6.07, *SE* = 2.67, *t*(436.32) = 2.28, *p* = .023; live vs mugshot video: *d* = 0.23 [0.07, 0.40], *b* = -5.26, *SE* = 2.56, *t*(436.27) = 2.05, *p* = .041). There were no significant differences in comfort ratings between the mugshot and full-body video conditions (*d* = 0.04 [-0.12, 0.20], *b* = -0.92, *SE* = 2.51, *t*(303.53) = 0.37, *p* = .714).

We also explored the relationship between comfort ratings and medium, choosing, and accuracy. The full model indicated a main effect in the live vs full-body video contrast (similar to the one already reported in the previous section) and a main effect of choosing: choosers reported higher levels of comfort (*M* = 77.22, *SD* = 21.84) than non-choosers (*M* = 70.57, *SD* = 25.13; *d* = 0.28 [0.15, 0.42], *b* = 7.14, *SE* = 1.21, *t*(571.60) = 5.92, *p* < 001). There were no other significant effects or interactions (highest *t* = 1.96, lowest *p* = .051).

Pre- and post-identification confidence. One indication of a live superiority hypothesis can be explored through self-reported confidence: if asked about the likelihood of success at a future identification, informing participants about what type of lineup will be presented could lead participants in the live condition to report higher confidence than participants in the non-live conditions. An analysis of variance showed a significant effect of the medium (*F*(2) = 5.27, *p* = .005),

although contrary to the predicted direction. Participants in the mugshot video condition reported highest confidence in their future identification (M = 62.15, SD = 22.00), followed by the full-body video condition (M = 58.95, SD = 20.61) and the live condition (M = 56.28, SD = 22.53). Tukey's post-hoc pairwise comparisons indicated that only the difference between live and mugshot video was significant (M_{Diff} = 5.86 [1.59, 10.14], p = .004). There were no significant differences in participants' ratings of identification confidence reported after they completed identification procedures for both (the male and the female) targets (live: M = 54.47, SD = 26.14; full-body: M = 54.80, SD = 26.16; $F(2)$ = 0.458, p = .633; mugshot: M = 56.41, SD = 25.61).

Endorsement of cues used to make identification decisions. Table SM6 displays rates of endorsement of cues used to make the identification decision across conditions. Compared to choosers in the mugshot condition, choosers in the live and full-body video conditions were less likely to endorse face cues and more likely to endorse height, posture, movement, and body cues. We explored associations between endorsed features and accuracy across conditions but found no significant results (highest z = 1.92, p = .055).

Table SM6

Cue endorsements for choosers across conditions

	Medium				
Feature	Live	Full-Body Video	Mugshot Video	χ^2	p
Face	83%	86%	95%	22.56	< .001
Height	49%	36%	3%	167.00	< .001
Posture	20%	17%	4%	41.94	< .001
Movement	12%	13%	6%	11.57	.003
Body	73%	67%	29%	141.57	< .001
Behavior	12%	11%	8%	2.61	.272

Endorsement of confidence-related statements. Table SM7 displays percentages of endorsement for additional indicators and justifications of choosing or not choosing a person from a lineup split by accuracy of identification decision. We also explored associations between these

indicators and accuracy and found that choosers were more likely to be accurate if they indicated that they would testify in court (*OR* = 3.37 [2.00, 5.77], $z = 4.51$, $p < .001$). There were no other significant associations.

Table SM7

Statements endorsed by choosers and non-choosers

Choosing	Endorsement	Accurate ID Decision			Inaccurate ID Decision		
		Live	Full-Body Video	Mugshot Video	Live	Full-Body Video	Mugshot Video
Chooser	Testify in court	26%	50%	27%	8%	19%	15%
	Selected closest looking person	57%	32%	60%	64%	55%	59%
	Just recognized	19%	20%	17%	20%	16%	23%
	Member looked like the person	7%	2%	2%	12%	10%	7%
	Would not identify again	2%	0%	0%	1%	3%	0%
Nonchooser	Sure person was not present	30%	19%	32%	22%	12%	27%
	Weak memory	11%	22%	13%	20%	18%	20%
	Wanted to first see others	18%	27%	16%	22%	30%	24%
	Saw someone who looked similar	28%	23%	27%	20%	30%	15%
	Would identify	8%	8%	9%	5%	12%	13%

Repeated lineup viewings. The sequential lineup procedure enabled participants to repeat a lineup if they explicitly requested it, or if they made identifications of multiple lineup members. Table SM8 displays the proportions of participants who repeated the lineup in the full-body video, mugshot video, and live conditions. Decisions made after repeated viewings were less accurate than decisions made after a single viewing of the lineup (23% vs. 52%) and Table SM9 indicates that after a repeated viewing, participants most frequently identified a filler.

Table SM8

Repeated viewings of lineups across conditions

Repetition	Medium					
	Live		Full-Body Video		Mugshot Video	
	Proportion	*n*	Proportion	*n*	Proportion	*n*
0	89%	227	81%	224	85%	277
1	10%	25	17%	47	13%	44
2	1%	3	< 1%	2	1%	4
3	-	0	< 1%	1	< 1%	1
4	-	0	< 1%	1	-	0

Table SM9

Final decisions after non-repeated and repeated viewings of lineups across conditions

	Target Present				Target Absent		
	Hit	Filler	Miss	*n*	Filler	Correct Reject	*n*
Non-repeated							
Live	.28	.22	.50	119	.29	.71	108
Full-Body Video	.40	.13	.47	109	.26	.74	115
Mugshot Video	.29	.21	.50	140	.27	.73	137
Repeated							
Live	.53	.47	.00	17	.82	.18	11
Full-Body Video	.22	.56	.22	27	.88	.12	24
Mugshot Video	.30	.48	.22	23	.88	.12	26

Not sure responses. Participants who did not make an identification could have indicated "not sure" for one or multiple lineup members. Table SM10 shows proportions of not sure decisions among non-choosers for the target and designated innocent suspect across conditions.

Table SM10

"Not Sure" decisions for the target and innocent suspect in rejected lineups across conditions

Lineup Member	Live	Full-Body Video	Mugshot Video	χ^2	*p*
Target	37%	49%	31%	4.75	.093
Innocent Suspect	19%	26%	14%	4.81	.090

Acknowledgements

We would like to thank the people who appeared as lineup members, the research assistants who made data collection possible (Sophie Berryman, Violet Burek, Simona Ciobotaru, Victoria Correia, Alexandra V. Costin, Vanessa Davis, Alice Durston, Jennifer Evans, Isabel Graham, Monica Gurung, Anežka Harasimová, Leigh Hollis, Kosmas Kakouris, Feni Kontogianni, Pamela Korsah, Marie Kubalová, Freya Lenton, Jacqueline Matthews, Georgia Maullin, Robin Miklica, Jagoda Mizera, Victoria Patching, Sharyne Patel, Christiana Petane, Joshua Quiza-Perez, Tristan Smith, Faith Snow, and Amelia Thomas), our colleagues who let us interrupt their lectures (Roger Moore, Paul Morris, Iris Nomikou, James Ost, Mark Turner, Julie Udell, Zarah Vernham), and the prospective students, parents, and undergraduate students who took part in the experiments. We would like to thank Tomáš Rubín and Matthew A. Palmer for their technical and programming support.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

Bergold, A. N., & Heaton, P. (2018). Does filler database size influence identification accuracy? *Law and Human Behavior, 42,* 227 -243.

Bruer, K. C., Fitzgerald, R. J., Price, H. L., & Sauer, J. D. (2017). How sure are you that this is the man you saw? Child witnesses can use confidence judgments to identify a target. *Law and Human Behavior, 41*, 541-555.

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

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29.

Cutler, B. L., & Fisher, R. P. (1990). Live lineups, videotaped lineups, and photoarrays. *Forensic Reports, 3*, 439-448.

Cutler, B. L., Fisher, R. P., & Chicvara, C. L. (1989). Eyewitness identification from live versus videotaped lineups. *Forensic Reports, 2*, 93-106.

Dent, H.R., Stephenson G.M. (1979). Identification evidence: Experimental investigations of factors affecting the reliability of juvenile and adult witnesses. In Farrington D.P., Hawkins K., Lloyd-Bostock S.M. (Eds.), *Psychology, law and legal processes* (pp. 196-206). Atlantic Highlands, NJ: Humanities Press.

Egan, D., Pittner, M., & Goldstein, A. G. (1977). Eyewitness identification: Photographs vs. live models. *Law and Human Behavior, 1*, 199-206.

Gleser, L. J. & Olkin, I. (2008). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357-377). New York: Russel Sage Foundation.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*, 3-10.

Kerstholt, J. H., Koster, E. R., & van Amelsvoort, A. G. (2004). Eyewitnesses: A comparison of live, video, and photo line-ups. *Journal of Police and Criminal Psychology, 19*, 15-22.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition, 5*, 21-33.

Lampinen, J. M., Smith, A. M., & Wells, G. L. (2019). Four utilities in eyewitness identification practice: Dissociations between receiver operating characteristic (ROC) analysis and expected utility analysis. *Law and Human Behavior, 43*, 26-44.

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.

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d′, not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition, 3*, 58-62.

Peters, D. P. (1991). The influence of stress and arousal on the child witness. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 60– 76). Washington, DC: American Psychological Association.

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Shepherd, J. W., Ellis, H. D., & Davies, G. M. (1982). *Identification evidence: A psychological evaluation*. Aberdeen: Aberdeen University Press.

Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., & Mackovichova, S. (2019). Deviation from Perfect Performance measures the diagnostic utility of eyewitness lineups but partial Area Under the ROC Curve does not. *Journal of Applied Research in Memory and Cognition, 8*, 50-59.

Sporer, S. L. (1991). Personenidentifizierungen bei Wahlgegenüberstellungen und Lichtbildvorlagen [Person identifications in live lineups and photospreads]. In R. Egg (Ed.), *Brennpunkte der rechtspsychologie: Polizei, justiz, drogen* (pp. 83-110). Bonn/Bad Godesberg: Forum.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702-712.

Tredoux, C., & Naylor, T. (2018). *r4lineups: Statistical inference on lineup fairness*. R package version 0.1.1. https://CRAN.R-project.org/package=r4lineups

Wells, G. L., & Penrod, S. D. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. In B. Rosenfield and S. D. Penrod (Eds.), *Research Methods in Forensic Psychology* (pp. 237-256). Hoboken, New Jersey: Wiley.

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition, 4*, 313-317.