A signal-detection analysis of eyewitness identification across the adult lifespan

Colloff, Melissa; Wade, Kimberley A.; Wixted, John T.; Maylor, Elizabeth A.

DOI:
10.1037/pag0000168

SUPPLEMENTAL MATERIALS

Excluded Carjacking and Theft Identification Responses

Although we initially planned to analyze the data from all four videos, in the end, we only analyzed the data from the *graffiti* and *mugging* videos because identification performance was very low for the other two videos, even for young subjects tested using fair lineups (*carjacking* $d' = 0.74$; *theft* $d' = 0.43$). For older subjects, performance was on the floor (*carjacking* $d' = –0.04$; *theft* $d' = –0.24$). Identification performance on the fair lineups was much better in the *graffiti* (young $d' = 1.21$, middle-aged $d' = 0.96$, older $d' = 0.65$) and *mugging* (young $d' = 1.08$, middle-aged $d' = 1.16$, older $d' = 0.70$) videos.

We report the identification responses made by each age group to the (included) *graffiti* and *mugging* videos for the replication, pixelation and block lineups in Table S4, and for the unfair do-nothing lineups in Table S6. Here, for completeness, we report the identification responses made by each age group to the (excluded) *carjacking* and *theft* videos for the replication, pixelation and block lineups in Table S1, and for the unfair do-nothing lineups in Table S2.

Table S1

Excluded Carjacking and Theft Identification Responses in Each Confidence Bin in the Replication, Pixelation and Block Lineups for the Young, Middle-aged and Older Adults

	Replication					Pixelation					Block				
	Target present			Target absent		Target present			Target absent		Target present			Target absent	
Confidence	Guilty Suspect	Foil	Incorrect Rejection	Foil	Correct Rejection	Guilty Suspect	Foil	Incorrect Rejection	Foil	Correct Rejection	Guilty Suspect	Foil	Incorrect Rejection	Foil	Correct Rejection
								Young							
0-20	0	3	4	1	2	0	5	2	8	0	3	4	4	4	5
30-40	3	1	4	8	5	4	5	6	12	4	3	8	5	4	3
50-60	6	6	14	11	12	5	20	9	17	11	5	10	4	11	19
70-80	11	7	7	12	9	1	10	8	8	12	4	6	10	13	12
90-100	1	7	12	6	15	4	4	13	5	12	4	4	9	2	15
								Middle-aged							
0-20	2	1	2	2	3	0	2	6	3	2	0	2	3	11	2
30-40	2	1	3	8	2	0	0	5	5	6	1	0	3	6	4
50-60	6	10	14	11	18	2	24	16	15	15	5	19	17	21	8
70-80	0	8	18	5	7	3	10	10	12	15	0	11	7	14	7
90-100	5	6	8	10	15	4	4	10	5	11	2	6	7	5	10
								Older							
0-20	0	3	4	3	4	0	3	2	7	2	1	5	3	5	4
30-40	0	2	8	5	5	1	12	5	7	7	1	6	5	5	3
50-60	7	13	15	21	12	1	18	20	16	22	3	13	14	23	15
70-80	1	13	8	13	8	2	11	12	8	7	0	12	11	12	11
90-100	0	6	6	6	4	0	5	4	4	9	2	4	3	5	5

Table S2

Excluded Carjacking and Theft Identification Responses in Each Confidence Bin in the Unfair Lineups for the Young, Middle-aged and Older Adults

	Target-present			Target-absent		
Confidence	Guilty Suspect	Foil	Incorrect Rejection	Innocent Suspect	Foil	Correct Rejection
			Young			
0-20	1	3	3	1	1	0
30-40	1	2	4	3	2	4
50-60	4	9	12	3	8	6
70-80	5	4	9	3	4	14
90-100	18	1	10	2	1	19
			Middle-aged			
0-20	1	2	2	2	1	4
30-40	1	6	4	1	4	4
50-60	8	7	10	3	5	9
70-80	9	2	8	3	6	8
90-100	13	1	12	3	5	13
			Older			
0-20	0	4	0	0	2	2
30-40	3	1	5	1	8	6
50-60	3	6	15	4	11	13
70-80	10	8	13	4	4	7
90-100	5	2	11	1	2	6

Background Performance Measures

We recruited the majority of our older adults from an organization that promotes lifelong learning. To check that we did not have an unusually able older adult sample, we examined whether our older adults showed the expected speed deficits in performance that accompany normal aging. Recall that our filler task consisted of three questionnaires followed by an anagram puzzle. Due to a technical error, there was missing filler task data for 25 older adults. The proportions of young, middle-aged and older adults who were still working on the questionnaire items at the end of the 8-min filler task were 0.10, 0.13, and 0.48, respectively. A 3 (age: young, middle-aged, older) × 2 (complete: yes, no) two-way chi-square analysis indicated that completion of the questionnaire items was dependent on age, χ^2 (2, N = 2,645) = 432.68, $p < .001$, Cramer's $V = .40$. Specifically, older adults were over 6 times more likely than the middle-aged adults, χ^2 (1, N = 1,755) = 256.16, $p < .001$, OR = 6.17, 95% CI [4.85, 7.89], and over 8 times more likely than the young adults, χ^2 (1, N = 1,755) = 307.05, $p < .001$, OR = 8.10, 95% CI [6.26, 10.57], to still be working on the questionnaire items at the end of the 8-min filler task. Young and middle-aged adults were equally likely to be working on the questionnaire items, χ^2 (1, N = 1,780) = 3.39, $p = .07$, OR = 1.31, 95% CI [0.97, 1.78].

Recall also that in the experimental task, subjects were asked to make an identification decision from a lineup and then rate their confidence in their decision. A one-way ANOVA showed that the length of time (s) to make an identification decision from the lineup was dependent on age group, F (2, 2667) = 43.69, $p < .001$. Older adults, M = 17.61, SD = 11.38, were slower than both middle-aged, M = 13.60, SD = 9.91, t (1745.3) = 7.94, $p < .001$, r = .19, and young adults, M = 13.63, SD = 9.90, t (1744.7) = 7.88, $p < .001$, r = .19, but middle-aged adults were not slower than young adults, t (1778) = 0.07, $p = .95$, r = .002. A second one-way ANOVA showed that the length of time (seconds) for subjects to provide a

confidence rating was also dependent on age group, F (2, 2667) = 45.40, p < .001. Older adults, M = 9.73, SD = 6.17, were slower than both middle-aged, M = 7.88, SD = 6.35, t (1776.5) = 6.23, p < .001, r = .15, and young adults, M = 7.13, SD = 5.16, t (1723.7) = 9.63, p < .001, r = .23. Middle-aged adults were also slower than young adults, t (1705.7) = 2.72, p = .007, r = .07. Together, these analyses suggest that our older adults showed the speed deficits in performance that accompany normal aging.

Preliminary Analyses: Performance on the Fair Lineups in the Graffiti and Mugging Videos

Before collapsing across the three fair lineup techniques (replication, pixelation and block) in our dataset, we checked that, within each age group, subjects performed similarly on the three fair lineup types. To this end, we examined subjects' identification responses, conducted ROC analysis and fit a signal-detection process model of identification performance to our data (Wixted & Mickes, 2014). To be clear, the following data and analyses refer to the included data from the graffiti and mugging videos.

Identification Responses

Figure S1 shows the identification responses made by the young, middle-aged and older adults in (A) target-present and (B) target-absent lineups, as a function of lineup type. Three 3 (lineup type: replication, pixelation, block) × 3 (identification response: guilty suspect, foil, rejection) two-way chi-square analyses indicated that performance was the same on the three fair lineups in the young, χ^2 (4, N = 688) = 2.25, p = .69, middle-aged, χ^2 (4, N = 688) = 1.90, p = .75, and older, χ^2 (4, N = 688) = 7.37, p = .12, adults.

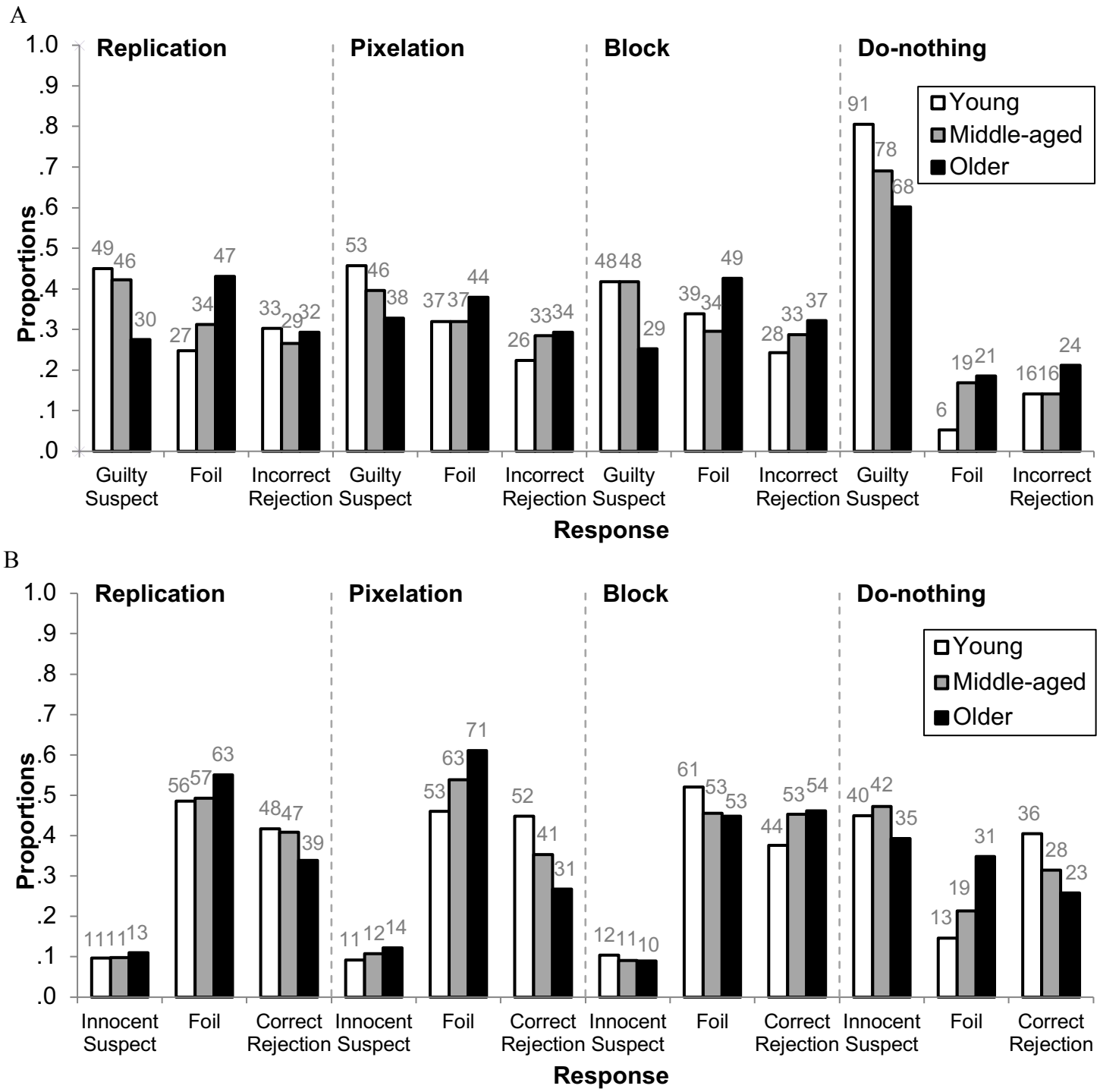

Figure S1. Identification responses made by the young, middle-aged and older adults in (A) target-present and (B) target-absent lineups, as a function of lineup type. Data are collapsed over the graffiti and mugging videos. In replication, pixelation and block target-absent lineups, the number of innocent suspect identifications was estimated by dividing the total number of false identifications by 6, and the number of foil identifications was estimated by dividing the total number of false identifications by 6 and then multiplying this by 5. In do-nothing target-absent lineups, the number of innocent suspect identifications was the number of times the person with the culprit's distinctive feature was identified and the number of foil identifications was the number of times a foil without the culprit's distinctive feature was identified. Data labels are absolute frequencies.

ROC Analysis

To confirm that ability to discriminate between innocent and guilty suspects was the same in the three fair lineups, we conducted ROC analysis. Figure S2 shows the ROC curves for the replication, pixelation, block and do-nothing (unfair) lineups for young, middle-aged, and older subjects. It is clear from Figure S2 that, within each age group, the ROCs for the replication, pixelation and block lineups lie on top of each other. This indicates that the three fair lineups led to similar levels of identification performance. The *p*AUCs statistically corroborated this (specificity = .910; see Table S3). In young adults, the *p*AUCs did not differ significantly between replication and pixelation ($D = 0.82$, $p = .41$), replication and block ($D = 0.69$, $p = .49$), or block and pixelation ($D = 0.18$, $p = .86$) lineups. Nor did the *p*AUCs differ significantly between replication and pixelation ($D = 0.42$, $p = .67$), replication and block ($D = 0.17$, $p = .86$), or block and pixelation ($D = 0.25$, $p = .80$) lineups in middle-aged adults. Finally, in the older adults, the *p*AUCs did not differ significantly between replication and pixelation ($D = 0.01$, $p = .99$), replication and block ($D = 0.46$, $p = .65$), or block and pixelation ($D = 0.44$, $p = .66$) lineups. Concordant with the analysis of identification responses, this suggests that, within each age group, all three fair techniques were equally effective at enhancing subjects' ability to discriminate between innocent and guilty suspects.

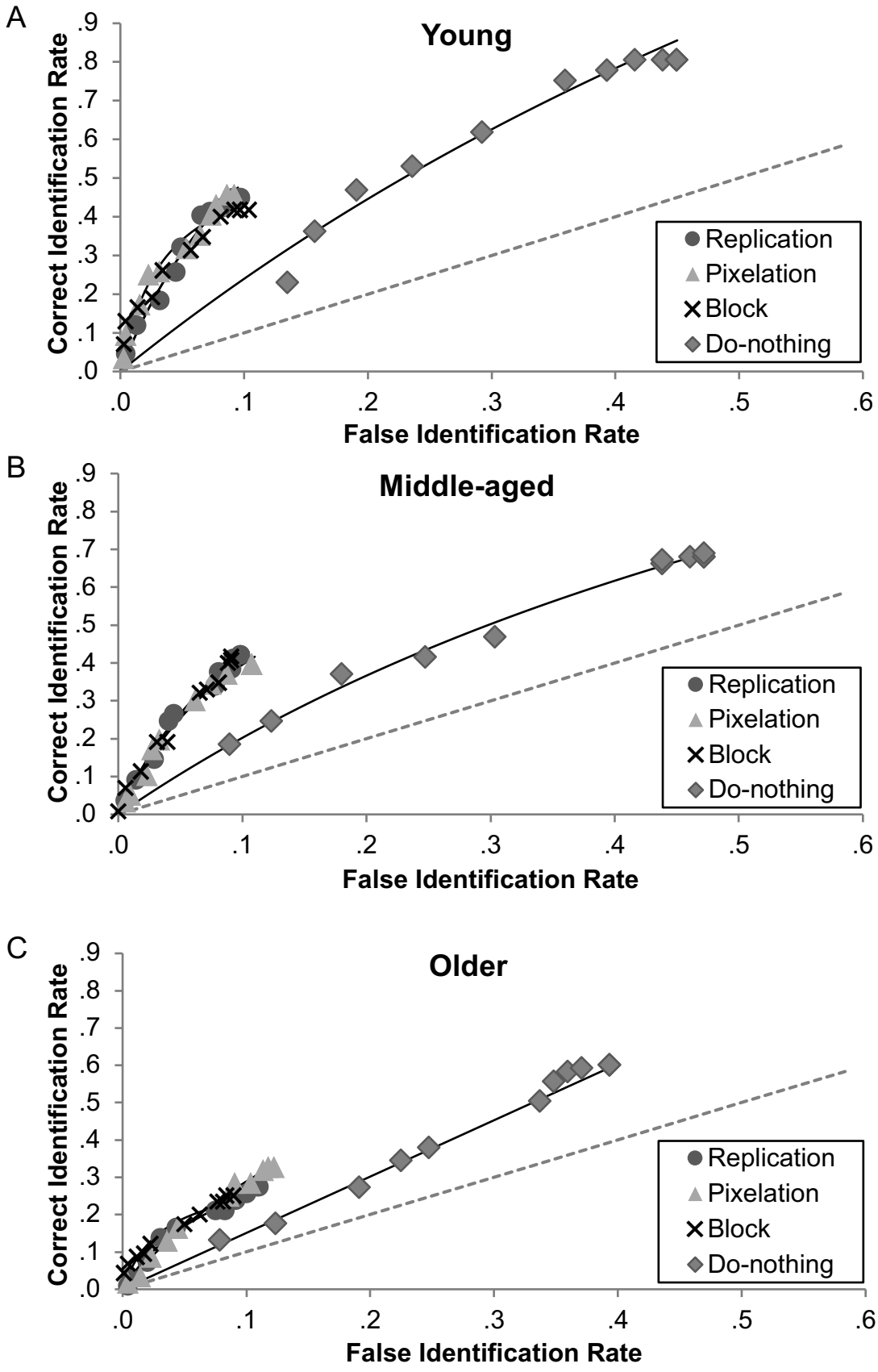

Figure S2. Receiver operating characteristic curves for the replication, pixelation, block and do-nothing (unfair) lineups for (A) young, (B) middle-aged, and (C) older adults. Data are collapsed over the graffiti and mugging videos. The dashed line represents chance-level performance.

Table S3
Partial Area Under the Curve (pAUC) Statistics [and 95% Confidence Intervals]

Lineup Type	Young	Middle-aged	Older
Replication	0.021 [0.010, 0.035]	0.018 [0.009, 0.032]	0.012 [0.005, 0.021]
Pixelation	0.028 [0.018, 0.040]	0.015 [0.007, 0.027]	0.012 [0.005, 0.022]
Block	0.027 [0.017, 0.037]	0.017 [0.008, 0.029]	0.014 [0.008, 0.023]

Note. Data are collapsed over the graffiti and mugging videos. Specificity (1 – FAR) = .910, which was set using the FAR range of the least extensive curve.

Modeling

To further assess whether all three fair lineup techniques were equally effective at enhancing ability to discriminate between innocent and guilty suspects, we fit a signal-detection process model to our data (Wixted & Mickes, 2014). We fit the model to the replication, pixelation and block data in each age group by minimizing the chi-square goodness-of-fit statistic. Within each age group, we performed three separate pairwise comparisons: replication versus pixelation, replication versus block, and pixelation versus block. We first fit the model allowing d' to differ across the two conditions being compared (unconstrained model). Table S4 shows our observed replication, pixelation and block data and the values predicted by the best-fitting signal-detection model for each age group, while Table S5 shows the best-fitting parameters and the chi-square goodness-of-fit statistics. It is clear from Table S4 that the model proficiently captured the trends in our data, and this is reflected in the chi-square goodness-of-fit statistics in Table S5.

To test whether there were any statistically significant differences in d' for each pairwise comparison, we fit the same model, allowing the confidence criteria to differ, but, critically, we constrained d' to be equal in the two conditions being compared. The overall χ^2, df and *p* rows in Table S5 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model did not provide a significantly worse

fit of the data for the replication and pixelation (young: χ^2 (1) = 0.18, p = .68; middle-aged: χ^2 (1) = 0.15, p = .70; older: χ^2 (1) = 0.02, p = .89), replication and block (young: χ^2 (1) = 0.01, p = .91; middle-aged: χ^2 (1) = 0.03, p = .86; older: χ^2 (1) = 0.14, p = .71) and pixelation and block (young: χ^2 (1) = 0.31, p = .58; middle-aged: χ^2 (1) = 0.35, p = .55; older: χ^2 (1) = 0.07, p = .79) comparisons. These results indicate that, within each age group, there was no statistically significant difference in d' between the three fair lineup conditions. Overall, our analyses suggest that performance was the same on the three fair lineup types. Therefore, for ease of interpretation, we collapsed the data over the replication, pixelation and block lineups within each age group.

Table S4

Observed and Predicted Identification Responses in Each Confidence Bin in the Replication, Pixelation and Block Lineups for the Young, Middle-aged and Older Adults

	Replication					Pixelation					Block				
	Target-present			Target-absent		Target-present			Target-absent		Target-present			Target-absent	
Confidence	Guilty Suspect	Foil	Incorrect Rejection	Foil	Correct Rejection	Guilty Suspect	Foil	Incorrect Rejection	Foil	Correct Rejection	Guilty Suspect	Foil	Incorrect Rejection	Foil	Correct Rejection
								Young							
0-20															
Observed	3.00	4.00	-	11.00	-	6.00	8.00	-	13.00	-	2.00	7.00	-	16.00	-
Predicted	3.27	4.49	-	9.73	-	5.30	6.83	-	14.80	-	4.54	6.61	-	14.09	-
30-40															
Observed	2.00	7.00	-	11.00	-	10.00	7.00	-	14.00	-	10.00	14.00	-	17.00	-
Predicted	4.26	5.32	-	10.69	-	7.26	8.04	-	15.44	-	8.90	11.06	-	20.77	-
50-60															
Observed	16.00	11.00	-	14.00	-	8.00	6.00	-	21.00	-	14.00	8.00	-	22.00	-
Predicted	10.48	10.95	-	19.63	-	10.52	9.43	-	15.93	-	12.25	11.85	-	19.26	-
70-80															
Observed	15.00	4.00	-	22.00	-	18.00	11.00	-	13.00	-	7.00	6.00	-	15.00	-
Predicted	14.26	10.49	-	16.22	-	16.69	10.14	-	14.79	-	11.00	7.56	-	10.84	-
90-100															
Observed	13.00	1.00	-	9.00	-	11.00	5.00	-	3.00	-	15.00	4.00	-	3.00	-
Predicted	12.92	4.65	-	6.33	-	12.06	3.30	-	4.21	-	12.17	4.39	-	5.70	-
Total															
Observed	-	-	33.00	-	48.00	-	-	26.00	-	52.00	-	-	28.00	-	44.00
Predicted	-	-	27.92	-	52.39	-	-	26.44	-	50.83	-	-	24.68	-	46.33
								Middle-aged							
0-20															
Observed	4.00	6.00	-	5.00	-	3.00	8.00	-	14.00	-	8.00	7.00	-	7.00	-
Predicted	2.83	4.32	-	9.01	-	4.33	7.06	-	13.34	-	4.61	6.52	-	12.82	-
30-40															
Observed	5.00	2.00	-	9.00	-	8.00	7.00	-	18.00	-	3.00	4.00	-	11.00	-
Predicted	3.20	4.55	-	8.91	-	6.66	9.63	-	16.59	-	3.83	4.91	-	8.96	-

50-60															
Observed	10.00	12.00	-	26.00	-	15.00	14.00	-	24.00	-	15.00	9.00	-	24.00	-
Predicted	10.80	13.03	-	22.95	-	13.21	15.29	-	23.32	-	12.38	13.18	-	21.67	-
70-80															
Observed	17.00	8.00	-	18.00	-	14.00	3.00	-	12.00	-	14.00	10.00	-	18.00	-
Predicted	13.25	11.52	-	17.59	-	10.18	8.35	-	11.33	-	14.96	10.72	-	15.27	-
90-100															
Observed	10.00	6.00	-	10.00	-	6.00	5.00	-	7.00	-	8.00	4.00	-	4.00	-
Predicted	12.59	5.72	-	7.75	-	9.08	4.15	-	5.21	-	9.13	3.15	-	4.03	-
Total															
Observed	-	-	29.00	-	47.00	-	-	33.00	-	41.00	-	-	33.00	-	53.00
Predicted	-	-	27.20	-	48.79	-	-	28.07	-	46.22	-	-	31.61	-	54.25
							Older								
0-20															
Observed	4.00	7.00	-	13.00	-	5.00	6.00	-	13.00	-	2.00	10.00	-	9.00	-
Predicted	3.40	7.63	-	12.73	-	3.45	7.72	-	12.59	-	3.46	6.92	-	11.19	-
30-40															
Observed	3.00	6.00	-	11.00	-	4.00	12.00	-	12.00	-	7.00	7.00	-	19.00	-
Predicted	3.06	6.44	-	10.26	-	4.40	9.13	-	14.07	-	5.87	10.73	-	16.37	-
50-60															
Observed	8.00	23.00	-	31.00	-	14.00	13.00	-	35.00	-	9.00	25.00	-	23.00	-
Predicted	11.33	20.53	-	30.32	-	11.71	20.76	-	29.40	-	12.54	18.68	-	26.11	-
70-80															
Observed	12.00	9.00	-	16.00	-	11.00	10.00	-	15.00	-	3.00	6.00	-	9.00	-
Predicted	9.01	12.06	-	16.30	-	8.43	11.48	-	14.91	-	5.23	5.80	-	7.51	-
90-100															
Observed	3.00	2.00	-	5.00	-	4.00	3.00	-	10.00	-	8.00	1.00	-	3.00	-
Predicted	3.51	2.90	-	3.73	-	6.10	5.41	-	6.66	-	5.02	3.71	-	4.63	-
Total															
Observed	-	-	32.00	-	39.00	-	-	34.00	-	31.00	-	-	37.00	-	54.00
Predicted	-	-	29.13	-	41.66	-	-	27.40	-	38.38	-	-	37.05	-	51.19

Note. Data are collapsed over the graffiti and mugging videos. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made.

Table S5

Full and Constrained Model Fits for the Replication vs. Pixelation, Replication vs. Block, and Pixelation vs. Block Comparisons in the Young, Middle-aged, and Older Adults

	Young				Middle-aged				Older			
	Full Model		Constrained Model		Full Model		Constrained Model		Full Model		Constrained Model	
Estimate	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation
μ_{guilty} (d')	1.18	1.25	1.20	1.20	1.09	1.02	1.05	1.05	0.70	0.72	0.71	0.71
c_1	1.16	1.13	1.17	1.13	1.11	1.07	1.11	1.08	1.01	0.96	1.01	0.96
c_2	1.30	1.34	1.30	1.33	1.24	1.25	1.23	1.26	1.19	1.14	1.19	1.13
c_3	1.45	1.57	1.46	1.57	1.36	1.49	1.36	1.50	1.33	1.33	1.33	1.33
c_4	1.80	1.89	1.81	1.88	1.74	1.95	1.74	1.96	1.86	1.83	1.86	1.83
c_5	2.35	2.50	2.42	2.55	2.27	2.43	2.27	2.43	2.54	2.33	2.55	2.33
Overall χ^2	24.93		25.11		19.69		19.83		18.75		18.77	
Overall df	18		19		18		19		18		19	
Overall p	.13		.16		.35		.40		.41		.47	
	Replication	Block	Replication	Block	Replication	Block	Replication	Block	Replication	Block	Replication	Block
μ_{guilty} (d')	1.18	1.16	1.17	1.17	1.09	1.12	1.10	1.10	0.70	0.77	0.74	0.74
c_1	1.16	1.07	1.16	1.07	1.11	1.17	1.11	1.17	1.01	1.13	1.02	1.13
c_2	1.30	1.26	1.29	1.26	1.24	1.35	1.24	1.35	1.19	1.28	1.19	1.28
c_3	1.45	1.56	1.45	1.57	1.36	1.48	1.36	1.48	1.33	1.52	1.34	1.52
c_4	1.80	1.96	1.80	1.96	1.74	1.89	1.75	1.88	1.86	2.09	1.87	2.09
c_5	2.35	2.40	2.35	2.40	2.27	2.52	2.27	2.52	2.54	2.47	2.55	2.46
Overall χ^2	29.01		29.02		18.61		18.64		18.14		18.28	
Overall df	18		19		18		19		18		19	
Overall p	.05		.07		.42		.48		.45		.50	
	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block
μ_{guilty} (d')	1.25	1.16	1.20	1.20	1.02	1.12	1.07	1.07	0.72	0.77	0.75	0.75
c_1	1.13	1.07	1.13	1.07	1.07	1.17	1.08	1.17	0.96	1.13	0.96	1.13
c_2	1.34	1.26	1.33	1.27	1.25	1.35	1.26	1.34	1.14	1.28	1.14	1.28
c_3	1.57	1.56	1.56	1.57	1.49	1.48	1.50	1.47	1.33	1.52	1.33	1.52
c_4	1.89	1.96	1.88	1.97	1.95	1.89	1.97	1.88	1.83	2.09	1.83	2.09
c_5	2.50	2.40	2.49	2.41	2.43	2.52	2.43	2.51	2.33	2.47	2.34	2.46
Overall χ^2	18.14		18.45		19.41		19.76		27.78		27.85	
Overall df	18		19		18		19		18		19	
Overall p	.45		.49		.37		.41		.07		.09	

Note. Data are collapsed over the graffiti and mugging videos. The full (unconstrained) model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

Modeling

To confirm our findings from the ROC analysis, we fit a signal-detection process model to our data (Wixted & Mickes, 2014). We fit the model to the fair lineups in each age group, and we discuss these findings in detail in the main text because they further our theoretical understanding of how identification behavior changes with age. Here, we present the model fits for the remaining comparisons examined in our ROC analysis. That is, we compare performance on the unfair lineups across age groups, and we compare performance on the fair and unfair lineups, within each age group.

The model for a fair lineup is described in detail in the main text, but the model for an unfair lineup differs slightly. Recall that in the unfair target-absent lineups, the innocent suspect had the culprit's distinctive feature but the other foils did not. For an unfair lineup such as this, the model consists of three distributions. Namely, the model assumes that the memory strength values (i.e., degree of familiarity) for guilty suspects, innocent suspects and foils have Gaussian distributions with means of μ_{guilty}, $\mu_{innocent}$, and μ_{foil}, respectively. The measure of interest is the distance between the μ_{guilty} and $\mu_{innocent}$ distributions (d'), which, similarly to the ROC analysis, reflects the ability to discriminate between guilty and innocent suspects.

The unfair lineup data contained 20 degrees of freedom, corresponding to the 5 levels of confidence for guilty suspect identifications and foil identifications in target-present lineups, and the 5 levels of confidence for innocent suspect identifications and foil identifications in target-absent lineups. Once these response frequencies were known, the number of rejections made in target-present and target-absent lineups was fixed. The model had 7 free parameters (μ_{guilty}, μ_{foil}, c_1, c_2, c_3, c_4, c_5) because we fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1, for simplicity. Thus, the fit had 20 – 7 = 13 degrees of freedom.

First, we examined how performance changed with age on the unfair lineups. We fit the model to the unfair lineup data in each age group by minimizing the chi-square goodness-of-fit statistic. Table S6 shows our observed unfair data and the values predicted by the best-fitting signal-detection model for each age group, whereas Table S7 shows the best-fitting parameters and the chi-square goodness-of-fit statistics. Again, it is clear from Table S6 that this simple model proficiently captured the trends in our data, and this is reflected in the chi-square goodness-of-fit statistics in Table S7.

Table S6
Observed and Predicted Identification Responses in Each Confidence Bin in the Unfair Lineups for the Young, Middle-aged and Older Adults

	Target present			Target absent		
Confidence	Guilty Suspect	Foil	Incorrect Rejection	Innocent Suspect	Foil	Correct Rejection
			Young			
0-20						
Observed	0.00	2.00	-	3.00	4.00	-
Predicted	3.79	1.35	-	4.05	2.53	-
30-40						
Observed	6.00	1.00	-	5.00	1.00	-
Predicted	5.08	1.48	-	4.78	2.56	-
50-60						
Observed	25.00	1.00	-	11.00	4.00	-
Predicted	19.13	3.54	-	13.79	5.20	-
70-80						
Observed	19.00	2.00	-	7.00	2.00	-
Predicted	16.35	1.50	-	8.31	1.81	-
90-100						
Observed	41.00	0.00	-	14.00	2.00	-
Predicted	42.62	1.02	-	11.32	1.02	-
Total						
Observed	-	-	16.00	-	-	36.00
Predicted	-	-	17.15	-	-	33.62
			Middle-aged			
0-20						
Observed	1.00	3.00	-	1.00	4.00	-
Predicted	2.83	1.90	-	2.74	2.76	-
30-40						
Observed	2.00	2.00	-	2.00	7.00	-
Predicted	4.80	2.79	-	4.27	3.82	-
50-60						

Observed	28.00	9.00	-	17.00	4.00	-
Predicted	23.31	8.77	-	16.47	10.49	-
70-80						
Observed	19.00	5.00	-	11.00	2.00	-
Predicted	18.89	3.47	-	9.77	3.50	-
90-100						
Observed	28.00	0.00	-	11.00	2.00	-
Predicted	28.05	1.60	-	9.14	1.42	-
Total						
Observed	-	-	16.00	-	-	28.00
Predicted	-	-	16.59	-	-	24.64
			Older			
0-20						
Observed	2.00	4.00	-	3.00	11.00	-
Predicted	6.31	5.25	-	5.28	6.10	-
30-40						
Observed	9.00	3.00	-	2.00	6.00	-
Predicted	6.61	4.53	-	5.04	4.95	-
50-60						
Observed	18.00	10.00	-	10.00	11.00	-
Predicted	18.45	8.94	-	12.14	9.02	-
70-80						
Observed	19.00	3.00	-	9.00	3.00	-
Predicted	16.34	4.49	-	8.76	4.11	-
90-100						
Observed	20.00	1.00	-	11.00	0.00	-
Predicted	19.93	2.03	-	7.70	1.71	-
Total						
Observed	-	-	24.00	-	-	23.00
Predicted	-	-	20.13	-	-	24.20

Note. Data are collapsed over the graffiti and mugging videos. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made.

To compare performance on the unfair lineups, we performed three separate pairwise comparisons: young versus middle-aged, young versus older, and middle-aged versus older. We fit the same model, allowing the confidence criteria to differ, but, critically, we constrained d' to be equal in the two age groups being compared. The overall χ^2, df and p rows in Table S7 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model did not provide a significantly worse fit of the data for the young and middle-aged adults, χ^2 (1) = 1.23, p = .27, the young and older adults, χ^2 (1) = 3.32, p = .07, or the middle-aged and older adults, χ^2 (1) = 0.53, p = .47. These

results support the ROC analysis and suggest that there was no statistically significant difference in *d'* between the three age groups on the unfair lineup. Therefore, regardless of which type of analysis we use, our conclusion remains the same: unfair lineups make it difficult for subjects of all age groups to discriminate between innocent and guilty suspects.

Table S7

Full and Constrained Model Fits for the Young vs. Middle-Aged, Young vs. Older, and Middle-aged vs. Older Unfair Lineup Comparisons

	Full Model		Constrained Model	
Estimate	Young	Middle-aged	Young	Middle-aged
μ_{guilty} (d')	0.83	0.59	0.70	0.70
$\mu_{innocent}$	0.00	0.00	0.00	0.00
μ_{foil}	–1.67	–1.44	–1.74	–1.38
c_1	–0.03	–0.12	–0.09	–0.06
c_2	0.12	0.00	0.05	0.06
c_3	0.28	0.17	0.21	0.23
c_4	0.76	0.78	0.69	0.84
c_5	1.14	1.26	1.06	1.33
Overall χ^2	33.05		34.28	
Overall df	26		27	
Overall p	.16		.16	
	Young	Older	Young	Older
μ_{guilty} (d')	0.83	0.43	0.62	0.62
$\mu_{innocent}$	0.00	0.00	0.00	0.00
μ_{foil}	–1.67	–1.29	–1.78	–1.19
c_1	–0.03	–0.05	–0.13	0.06
c_2	0.12	0.19	0.01	0.29
c_3	0.28	0.39	0.16	0.49
c_4	0.76	0.88	0.64	0.98
c_5	1.14	1.36	1.01	1.45
Overall χ^2	32.95		36.27	
Overall df	26		27	
Overall p	.16		.11	
	Middle-aged	Older	Middle-aged	Older
μ_{guilty} (d')	0.59	0.43	0.51	0.51
$\mu_{innocent}$	0.00	0.00	0.00	0.00
μ_{foil}	–1.44	–1.29	–1.48	–1.25
c_1	–0.12	–0.05	–0.16	0.00
c_2	0.00	0.19	–0.04	0.23
c_3	0.17	0.39	0.13	0.43
c_4	0.78	0.88	0.74	0.92
c_5	1.26	1.36	1.22	1.40
Overall χ^2	36.60		37.13	
Overall df	26		27	
Overall p	.08		.09	

Note. Data are collapsed over the graffiti and mugging videos. The full model allows d' to differ between the two age groups being compared. The constrained model holds d' constant across the two age groups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two age groups together.

Next, we compared performance on the fair and unfair lineups within each age group. The observed and predicted data for the fair lineups are shown in Table 3 in the main text, whereas the observed and predicted data for the unfair lineups are shown in Table S6. We performed a separate fair versus unfair pairwise comparison for the young, middle-aged and older adults. Again, we first fit the model allowing d' to differ across the fair and unfair conditions. We then fit the same model, allowing the confidence criteria to differ, but constraining d' to be equal across the fair and unfair lineups. The overall χ^2, df and p rows in Table S8 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model provided a significantly worse fit of the data for the young adults, χ^2 (1) = 5.13, p = .02, and middle-aged adults, χ^2 (1) = 8.67, p = .003, but this did not reach statistical significance in the older adults, χ^2 (1) = 3.03, p = .08. These results indicate that there was a statistically significant difference in d' between the fair and unfair lineups in the young and middle-aged adults, but this only approached significance in the older adults. Nevertheless, descriptively speaking, these results are consistent with our ROC analyses. Both analyses suggest that all three age groups were less able to distinguish between innocent and guilty suspects in the unfair lineups than in the fair lineups.

Table S8

Full and Constrained Model Fits for the Fair vs. Unfair Lineup Comparisons in the Young, Middle-aged, and Older Adults

	Young				Middle-aged				Older			
	Full Model		Constrained Model		Full Model		Constrained Model		Full Model		Constrained Model	
Estimate	Fair	Unfair	Fair	Unfair	Fair	Unfair	Fair	Unfair	Fair	Unfair	Fair	Unfair
μ_{guilty} (d')	1.21	0.83	1.14	1.14	1.07	0.59	0.98	0.98	0.72	0.43	0.66	0.66
μ_{foil}	-	–1.67	-	–1.50	-	–1.44	-	–1.23	-	–1.29	-	–1.16
c_1	1.13	–0.03	1.12	0.14	1.12	–0.12	1.11	0.08	1.04	–0.05	1.03	0.08
c_2	1.31	0.12	1.29	0.28	1.28	0.00	1.26	0.20	1.21	0.19	1.20	0.31
c_3	1.54	0.28	1.53	0.45	1.44	0.17	1.43	0.36	1.40	0.39	1.39	0.51
c_4	1.89	0.76	1.88	0.94	1.86	0.78	1.84	0.99	1.92	0.88	1.91	1.00
c_5	2.44	1.14	2.42	1.31	2.39	1.26	2.38	1.47	2.45	1.36	2.44	1.47
Overall χ^2	23.72		28.85		35.44		44.11		25.13		28.16	
Overall df	22		23		22		23		22		23	
Overall p	.36		.19		.03		.01		.29		.21	

Note. Data are collapsed over the graffiti and mugging videos. The full model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

Confidence and Accuracy in Young-old and Old-old Adults

Older adults made slightly (but not significantly) less accurate suspect identifications at every level of confidence than did young and middle-aged adults. To investigate this further, we separated our older adults into young-old (aged 60–70, $n = 463$) and old-old (aged 71+, $n = 225$) groups. We used the same method we report in the main paper to plot suspect identification accuracy as a function of confidence. Figure S3 shows the confidence-accuracy curves for the fair lineups in the young-old and old-old groups. First, it is important to note that the unexpectedly poor accuracy of the old-old adults at the highest level of confidence should be treated with caution because there were only four subjects in this age group who made an identification decision with this level of confidence. Focusing on the remaining confidence levels (i.e., 0-20, 30-40, 50-60, 70-80), we can see that, as before, the error bars for each age group overlap. This indicates that the differences in suspect identification accuracy between the age groups at each level of confidence are not statistically reliable. Nevertheless, the same numerical trend that we observed in our main confidence-accuracy analysis is apparent: old-old adults are slightly (but not significantly) less accurate at every level of confidence than the young-old adults. This suggests that as memory ability declines with age, older adults do not adjust their criteria to the extent required for them to be just as accurate at each level of confidence as their younger counterparts.

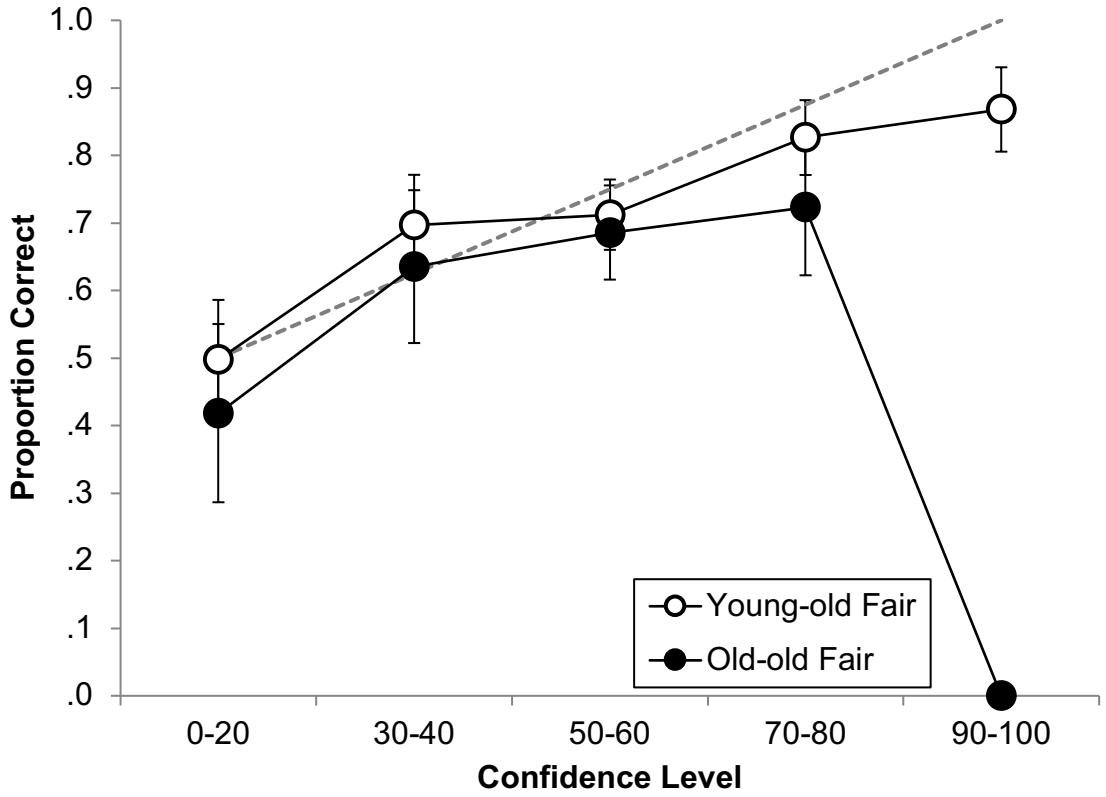

Figure S3. Confidence-accuracy curves for suspect identifications made by young-old and old-old adults in the fair lineups. Error bars ±1 *SE*. Data are collapsed over the graffiti and mugging videos. The dashed diagonal line signifies chance-level accuracy at the lowest confidence bin (0-20) and perfect accuracy at the highest confidence bin (90-100).