Applied Cognitive Psychology

WILEY

RESEARCH ARTICLE OPEN ACCESS

Seeing Faces Differently: Assessing the Influence of Children's Perceived Similarity on Eyewitness Identification Accuracy

Kaila C. Bruer[1] | Heather L. Price[2]

[1]Luther College at the University of Regina, Regina, Canada | [2]Thompson Rivers University, Kamloops, Canada

Correspondence: Kaila C. Bruer (kaila.bruer@uregina.ca)

Received: 19 June 2025 | **Revised:** 18 November 2025 | **Accepted:** 19 November 2025

Keywords: adults | children | eyewitness | filler selection | lineup stimuli construction

ABSTRACT

We investigated whether using children's perceived similarity to construct lineups changed children's identification performance. After a pilot showing that children rate suspects and fillers as more similar than adults do, we ran three experiments with child eyewitnesses (ages 6–11, younger group 6–8; older, 9–11) and an adult comparison group (18- to 58-years-old; Experiment 3). We analyzed accuracy, discriminability, confidence-accuracy calibration, and decision patterns (suspect identification, filler identification, or rejection), as a function of both target presence and lineup creator (adult- or child-created). Experiments 1 and 2 found that child-created lineups improved children's pattern of responding to suggest better discriminability and better confidence-accuracy calibration. In Experiment 3, we extended the design to include an adult sample, finding no effect of lineup type for adult witnesses and limited benefits for children. These results suggest that age-matched similarity information can improve children's lineup performance under some conditions, but the benefits are not universal.

1 | Introduction

Over the last few decades, developmental lineup researchers have focused on understanding a well-documented problem: younger children, when compared to older children and adults, struggle to make correct decisions from lineups that do not contain the actual perpetrator in the lineup. That is, children are more likely to falsely identify an innocent person when the actual perpetrator is not present (see Fitzgerald and Price 2015).

Researchers have theorized that a combination of social (e.g., Pozzulo et al. 2012) and cognitive elements (e.g., Charman and Wells 2007; Dunlevy and Cherryman 2013; Hiller and Weber 2013; Pozzulo and Lindsay 1999) likely drives children's propensity to identify an innocent person. For instance, Hiller and Weber (2013) highlighted how, when compared to other recognition memory tasks, children appear to uniquely struggle with eyewitness identification: "...the substantially higher false identification rates observed for child, over adult, witnesses appear likely to be a function of the identification context itself, not a consequence of basic cognitive or metacognitive development" (p. 190). So, what is it about the identification context that contributes to children's tendency to over choose from perpetrator-absent arrays? In a series of three experiments, we examined whether the process of selecting fillers to include in a lineup contributes to children's over choosing.

1.1 | Filler Selection

An eyewitness lineup typically comprises a suspect (guilty or innocent) and fillers (known innocents). An important part of constructing a lineup for an eyewitness is ensuring that the lineup is fair and that the suspect does not "unduly stand out" (Technical Working Group for Eyewitness Evidence, 2003, p. 32). There are two primary methods for selecting fillers when building a lineup. The first is based on the perceived similarity to the suspect (i.e., the match-to-appearance approach; Police Executive Research Forum 2013). This is the method most used by police (Wogalter et al. 2004). Judgments of perceived similarity are typically made either by an investigator/experimenter directly involved in the investigation or experimental design (e.g., Beresford and Blades 2006; Gross and Hayne 1996) or by others (e.g., research participants) who are independent from the lineup task (e.g., Havard and Memon 2013; Karageorge and Zajac 2011; Marin et al. 1979). An alternative method for building a lineup is to use the description-match strategy in which fillers are selected based on a witness's description of the culprit, rather than on the suspect's physical appearance (Wogalter et al. 1992).

There has been discussion in the literature about which is the more effective method to select fillers (e.g., Lindsay and Wells 1980 for match-to-appearance; description-match method: Clark 2003; Luus and Wells 1991; Malpass et al. 2007; Wells et al. 1993). Regardless of the methods, most research has relied on adult perceptions to select fillers. In the most recent meta-analysis examining eyewitness performance across the lifespan, 64 research papers were reviewed that included children as eyewitness participants (Fitzgerald and Price 2015). Of those 64 studies, 37[1] discussed the procedures used when choosing fillers and innocent suspects for their lineups. Of these, many ($n=24$) researchers described using similarity ratings to select fillers and, thus, used the match-to-appearance strategy (e.g., Gross and Hayne 1996; Fitzgerald et al. 2014; Pozzulo and Dempsey 2006; Pozzulo et al. 2012; Parker and Ryan 1993), while others ($n=10$) used a description of the perpetrator to select fillers (e.g., Beal et al. 1995; Beresford and Blades 2006; Brewer and Day 2005; Dehon et al. 2013; Dunlevy and Cherryman 2013).[2]

Of the 37 research papers reviewed that all those included in the review are found in the reference page and are indicated with an asterisk (*), all but two (Havard et al. 2010, 2012) created lineup stimuli based on the judgments of adults. In 2010, Havard and colleagues examined lineup performance of children (aged 7–9) and adolescents (aged 13–15). Though the lineups were created by investigators, Havard and colleagues asked 14 children (mean age of 10, 8–11 range) to rate faces on a 7-point scale for distinctiveness (i.e., 'if you had to pick this person out of a crowd at a railway station, how easy would it be?'). Similarly, Havard et al. (2012) conducted this same check of their stimuli using two groups of raters: 12 children (mean age of 8.25, range 6–9) and 19 adults and found no significant differences in mean distinctiveness ratings. However, it is important to consider that distinctiveness ratings may differ from similarity ratings, which is the dimension on which lineup members are typically measured.

Overall, most of our conclusions about child eyewitness performance have been made using stimuli created with adult perceptions of filler-suspect similarity. As discussed by Malpass et al. (2007), there are likely benefits to ensuring that whoever makes the judgments about filler similarity is similar to the witness(es) in terms of demographic characteristics, including age. Perhaps then, having children make identification decisions from stimuli built using adult perceptions contributes to an 'identification context' for which children struggle.

1.2 | Developmental Differences in Facial Processing

There is reason to expect that children will process and perceive similarity between faces differently from adults. A major concept supporting developmental differences in similarity judgments is the development of facial expertise (see Bruce and Young 1998). Research typically examines the proficiency of facial recognition by separating featural (e.g., shape of eyes) and configural (e.g., spacing of the eyes) information processing. Configural processing involves the ability to perceive spatial relationships between features in a more holistic way while processing faces—encompassing both first-order relations (the canonical arrangement, such as two eyes positioned over a mouth), second-order relations (metric spacing between features), and holistic integration of features into a unified representation of the face (Maurer et al. 2002). Although face processing skill exists from birth (Mondloch et al. 1999) and children and adults both draw on featural cues and multiple forms of configural cues, there is considerable evidence to suggest that the relative weight of these cues changes with age and associated demands of the task (Maurer et al. 2002; for example, whether the face is familiar). Children aged 4 to 14 years rely more on featural processing to process unfamiliar faces, whereas configural processing of faces continues to develop well into adulthood (Gao et al. 2011; Mondloch et al. 2002; Taylor et al. 2001).

In considering that children rely relatively more on featural cues and adults more on holistic/configural use (first and second order relations), the face-space offers a unifying account of perceived similarity. In face-space, identities (faces) occupy locations in a multidimensional geometry; distinctness indexes distances from the norm and typically reflects local density (Valentine 1991). Because within-person variability clusters multiple images of the same identity more tightly than images of different identities (Jenkins et al. 2011), observers' pairwise similarity ratings approximate local distances in this shared space. Critically, if children and adults differentially weight featural versus configural/holistic information (Maurer et al. 2002; Searcy and Bartlett 1996; Want et al. 2003), then the perceived distances and local densities around a target will systematically differ by age.

Under this perspective, lineup confusability follows from how the target and fillers are arranged: when similar faces cluster around the target (high local density), errors rise; when the target is more distinctive (greater distance), errors fall (Valentine 1991). Consequently, the same lineup can differ in difficulty across age groups, and seemingly small filler-selection choices that alter local density can shift error patterns (Malpass et al. 2007; see Fitzgerald et al. 2014). This framework motivates testing age-matched similarity judgments for lineup construction: if children's and adults' similarity spaces differ, then child-derived versus adult-derived filler choices should produce measurably different lineup geometries—and thus different identification outcomes—especially for child witnesses. It is conceivable, then, that when asked to rate the similarity between two faces, children may focus on different aspects of a face than adults. This, in turn, may produce differences in judgments about how similar two (or more) faces are. Further evidence that children may perceive faces differently from adults can be found in the 'own-age' bias literature; children and younger adults exhibit better discriminability for same-age than different-age faces (see Rhodes and Anastasi 2012). Together, these findings indicate that children's perceptions of similarity may be systematically different from those of adults, and thus lineup construction based on adult perceptions could inadvertently disadvantage younger witnesses. Suppose a lineup was created using adult similarity ratings. In that case, children may find faces more similar (more challenging to differentiate), thus widening the gap between child and adult eyewitness identification accuracy. One approach to investigating this possibility is to look for differences in similarity ratings made by children and the adult demographic typically used in eyewitness research—undergraduate students.

Though we do not know the extent to which children evaluate facial similarity differently from adults, variations in adult-established similarity between the suspect and fillers have been shown to impact children's performance on a lineup task. In an examination of the impact of similarity ratings on child eyewitnesses, Fitzgerald et al. (2014) found that an increase in similarity among lineup members reduced identifications of both guilty and innocent suspects. Of note, Fitzgerald et al. (2014) found that increases in similarity reduced children's guilty suspect identification rates by 26% (moderate similarity = 0.74 vs. high = 0.48) but did not impact adult witnesses' guilty suspect identification rates (moderate = 0.76 vs. high = 0.74). It is possible that performance differences as a function of lineup similarity are magnified in the eyewitness literature for children because they perceive the similarity between members differently than the adults whose judgments were used to build the lineups.

1.3 | Present Research

We explored whether creating lineups based on children's perceptions of similarity would influence children's (Experiments 1–3) and adults' (Experiment 3) responses during a lineup task. Should such an influence be observed, it has implications for researchers who are interested in understanding why children often perform worse than adults on lineups that do not contain a match to memory (i.e., perpetrator-absent).

Our first step was to conduct a pilot study to examine whether children's perceptions of similarity between lineup members paralleled similarity ratings provided by adults. The findings supported perceived differences and, as such, we ran a series of three experiments, each with two suspects (i.e., target faces). In Experiment 1, we compared children's lineup identification decisions from lineups created with filler-suspect similarity ratings made either by adults or by children. Experiment 1 explored this question under relatively poor encoding conditions (brief culprit/target exposure via video). Experiment 2 explored the same question using different targets that were viewed under improved encoding conditions (10-min exposure to live suspects). In Experiment 3, we explored how using child-created (versus adult-created) stimuli influenced both child and adult eyewitnesses. We examined children's performance across two age groups: 6–8 years and 9–11 years. These groupings reflect developmental transitions in facial processing and decision-making. Younger school-age children (6–8) rely more heavily on featural and external cues, whereas older children (9–11) increasingly incorporate configural information in ways that approximate adult processing (Gao et al. 2011; Mondloch et al. 2002). Prior work has also shown systematic differences in lineup responding across these age ranges (e.g., Fitzgerald and Price 2015). Identifying differences (if any) in children's performance as a function of adult-created versus child-created lineups will allow for the facilitation of eliciting the best evidence from children and a more accurate interpretation of past comparisons between child and adult eyewitness identification performance.

1.3.1 | Experimental Overview

Table 1 outlines the three experiments (plus an adult comparison in Experiment 3) that used the same lineup procedure and outcome metrics while varying one element per Experiment. Experiment 1 established a child baseline (ages 6–8 vs. 9–11) with child- vs. adult-created lineups; Experiment 2 tested robustness in an independent child cohort with the same manipulation; Experiment 3 replicated the child design and added an

TABLE 1 | Experiment methods summary table.

Study	Number of suspects	Encoding type	Lineup source	Child age group	Similarity ratings sample size	Eyewitness sample size
Pilot	2	—	Adult-created	6–14 years	99 child, 98 adult	—
Exp. 1	2	Video	Adult- vs. Child-created	6–11 years	22 child, 24 adult	255 children
Exp. 2	2	Live	Adult- vs. Child-created	6–11 years	20 child, 18 adult	249 children
Exp. 3	2	Video	Adult- vs. Child-created	8–12 years	15 child, 16 adult	144 children, 142 adults

Note: The pilot focused only on perceptual similarity and, as such, had no encoding phase.

adult sample tested with the identical materials. Study-specific deviations from the shared protocol are noted within each subsection.

2 | Pilot

As a preliminary exploration, we examined whether child eyewitnesses perceive face similarity differently from adult eyewitnesses. Using lineup stimuli from a previously conducted study (Bruer and Price 2017), we recruited a new sample of children and adults. We asked participants to rate how similar each lineup member (fillers and innocent suspects) was to the associated suspect. We then conducted an initial exploration into the relationship between rated similarity and identification accuracy.

3 | Pilot Method

3.1 | Participants

We recruited 99 children, 6- to 14-years-old ($M_{\text{age}} = 9.22$, SD = 1.64; 62% female), from camps and after-school programs and 98 adults ($M_{\text{age}} = 22.00$, SD = 5.79; 86% female) from a university undergraduate participant pool. Participants were recruited from mostly affluent families in central Canada from a primarily white population (community ethnicity: white 79%; Indigenous 10%; Southeast Asian 3%; South Asian 3%). Participants were asked to rate the similarity of two culprits (male and female) to all members of their corresponding lineups. The wide age range was included due to the exploratory nature of this pilot and the convenience sample available to researchers.

3.2 | Materials

3.2.1 | Adult-Created Lineup Stimuli

We used two lineups (one male and one female lineup, both Caucasian in their early 20s) from a previously conducted eyewitness identification study with children (Fitzgerald et al. 2013) in which lineup fillers were selected from the Glasgow Unfamiliar Face Database (Burton et al. 2010). Like most experiments reviewed above, the original lineups were built using independent adults' ratings of similarity. In that study, independent adult (undergraduate students) judges[3] ($n = 24$) provided pairwise (subjective) similarity ratings between photographs of the suspect (i.e., target) and 100 potential fillers (i.e., preselected on gender) on a 10-point Likert-type scale (1 = not at all similar, 10 = highly similar). Mean ratings were used to select fillers, ensuring fillers were neither too similar nor too distinct from the suspects (Fitzgerald et al. 2013). The mean similarity rating for the female fillers was slightly lower ($M = 3.16$) than for the male ($M = 3.60$). Innocent suspects were selected as moderately similar-looking for the female lineup ($M = 3.32$) and the most similar-looking for the male lineup ($M = 5.86$). All lineup images (180H × 288W pixels) were displayed simultaneously on an 11-in. touchscreen tablet and shown using *Eprime 2.0* software that recorded participants' responses. See Figure 1 for adult-created lineup stimuli. The order of face presentation during the similarity ratings task was randomized for each participant to minimize order effects and potential response biases.

3.3 | Procedure

Both child assent and parent consent were obtained. Participants (children and adults) were shown lineup stimuli for one suspect

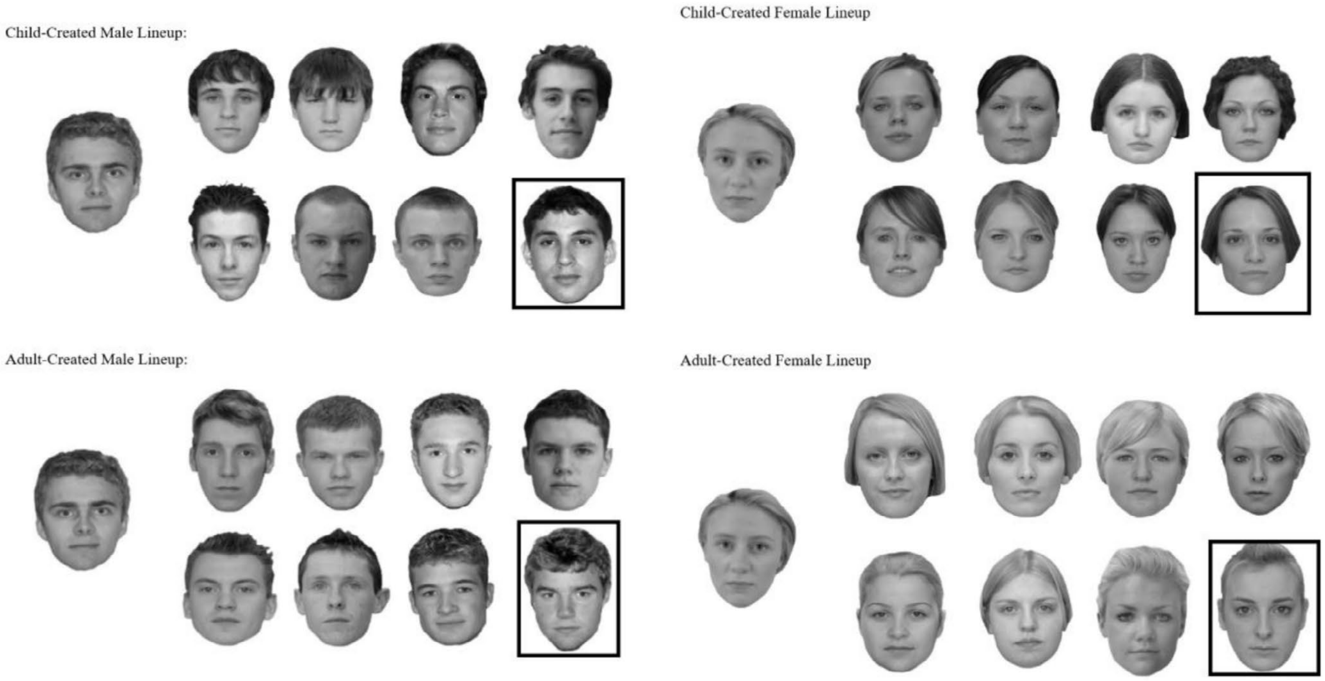
Child-Created Male Lineup:
Child-Created Female Lineup
Adult-Created Male Lineup:
Adult-Created Female Lineup

FIGURE 1 | Pilot and experiment 1 lineup stimuli. *Note:* Target is on the left and target replacement is indicated with the black box. Adult-created lineup was used in both Pilot and Experiment 1, while child-created lineups were only used in Experiment 1.

at a time, with presentation order counterbalanced. Participants were shown the first suspect's picture alongside each of the seven fillers and the innocent suspect, one pairing at a time, for a total of eight comparisons for each suspect. Each time a pair of images was shown, participants were asked to rate the similarity between the suspect and the other image. Below the two pictures was a 10-point similarity scale (1 = not at all similar, 10 = highly similar). Note that this similarity scale was not adapted to reflect children's development; instead, we used a similar approach to what is typically asked of adults. A research assistant read the following instructions prior to beginning the task:

> I am going to show you some pictures. One on this side and one on this side (*indicate side*). You will always see this man's picture on the left side of the screen—right here. His name is Matthew. On the other side of the screen you will see a picture of another man. Each time you see a new man, I want you to carefully look at each picture and tell me how much Matthew looks like the man. If Matthew looks a lot like the man, choose a big number (*indicate on the scale*). If Matthew doesn't look very much like the man, choose a small number. If Matthew looks a little bit like the man, pick a number somewhere in the middle.

After the task was complete for the first suspect, it was repeated for the second suspect. The entire procedure took less than 10 min.

4 | Pilot Results and Discussion

4.1 | Age Differences in Similarity Ratings

Mean similarity ratings provided by children and adults of the two suspects and their respective lineup members were examined. A 2 (Age: child, adult) × 2 (Suspect: Male, Female) analysis of variance (ANOVA) was conducted with average similarity ratings as the dependent variable. A main effect of age was found, $F(1, 369) = 6.28$, $p = 0.01$, $\eta^2 = 0.02$, such that children rated the fillers as significantly more similar to the suspects than did adults. This pattern was consistent across both suspects. There was also a main effect of suspect, $F(1, 369) = 3.78$, $p = 0.05$, $\eta^2 = 0.01$; the female suspect was rated as more similar to the fillers than was the male suspect. There was no interaction between age and suspect. See Table 2 for mean similarity ratings.

TABLE 2 | Mean similarity ratings (standard deviations) provided for each suspect.

	Female suspect	Male suspect	Total
Child	3.65[b] (1.51)	3.19[b] (1.35)	3.42[a] (1.45)
Adult	3.10 (1.51)	2.95 (1.62)	3.02[as] (1.56)
Total	3.38 (1.54)	3.07 (1.50)	3.22 (1.52)

[a] Indicates significant differences per columns.
[b] Indicates significant differences per row.

We also examined whether there were age-related differences in similarity ratings within the child sample. We split children into younger (6–8 years; $n = 59$) and older (9–11 years; $n = 126$) to match Experiments 1–3. Mean perceived similarity did not differ by age: 6–8, $M = 3.55$, SD = 1.52; 9–11, $M = 3.28$, SD = 1.28; Welch's $t(97.76) = 1.19$, $p = 0.237$, Hedges' $g = 0.19$. In contrast, ratings were more dispersed for younger children, indexed by both within-rater standard deviation (6–8, $M = 2.23$, SD = 0.98; 9–11, $M = 1.77$, SD = 0.73; Welch's $t[89.29] = 3.26$, $p = 0.002$, $g = 0.54$) and variance (6–8, $M = 5.743$, SD = 5.01; 9–11, $M = 3.65$, SD = 2.94; Welch's $t[72.36] = 2.92$, $p = 0.005$, $g = 0.51$). These exploratory results suggest greater variability in scale use (or noisier similarity judgments) among younger children but not a difference in central tendency. The pilot revealed that children rated similarity between faces differently from adults. The next step was to run a full experiment to see if this perceived similarity difference translated into differences in responding on a lineup.

5 | Experiment 1

We examined whether children's performance on a lineup task was influenced by whether children's or adults' similarity ratings were used to select lineup fillers. The experiment was conducted in two phases. First, we recruited an independent group of children to rate the similarity between the male and female suspects with 100 other faces (the same faces used with adults during the construction of lineups used during the pilot). Once these similarity ratings were gathered, we built new lineups using only children's ratings. In the second phase, we examined child eyewitness performance using these new, child-created lineup stimuli, relative to stimuli that were created using adult ratings (lineup stimuli from the Pilot). We hypothesized that child eyewitnesses would make more accurate lineup decisions when responding to lineup stimuli that were created with children's similarity ratings (i.e., child-created lineup) than when presented with stimuli created with adults' similarity ratings (i.e., adult-created lineup). We also expected older children (9–11) to outperform younger children (6–8), especially in target-present lineups. Both groups were predicted to benefit from child-created lineups, with younger children showing the greatest relative improvement.

6 | Experiment 1 Method

6.1 | Participants and Design

We recruited 22 children from the community to provide similarity ratings (7- to 12-years-old; $M_{age} = 9.09$, SD = 1.64; 48% female). We also recruited an additional 255 children, 6- to 11-years-old ($M_{age} = 8.71$, SD = 1.43; 41% female) from a local camp as eyewitnesses. These children were assigned to a 2 (Lineup-creator: Child, Adult) × 2 (Actor: Male, Female) × 2 (Age: 6–8, 9–11) × 2 (Target Presence: Present, Absent) mixed design (with Lineup-creator and Age as between-subjects variables) in which the presence of each suspect was randomly assigned by a computer for each participant. Children came from the same population as described in the Pilot Study, and research ethics board approval was obtained.

A post hoc sensitivity analysis (80% power, $\alpha = 0.05$) indicated that the minimum detectable Target Presence × Lineup-Creator interaction on accuracy was 1.18 log-odds (OR = 3.25; $n = 508$ trials). For the three-way interaction with Age, the MDES was 2.55 log-odds (OR = 12.85), indicating limited sensitivity to higher-order effects. For full sensitivity analysis results, see Supporting Information.

6.2 | Materials

6.2.1 | Adult-Created Lineup Stimuli

The adult-created stimuli were the same as discussed in the Pilot.

6.2.2 | Child-Created Lineup Stimuli

To construct the child-created lineup, we followed the same procedure used for the adult-created lineup. Children ($n = 22$) visited the lab and were paired with one of three female research assistants. Once assent was received, children were shown the same 200 photographs (100 females, 100 males) used in the pilot. Instructions (see pilot), including an example image, were verbally delivered, and images were shown via a touchscreen tablet. Children made their selection by touching their selection on the tablet. After completing the ratings for the first suspect, children took a 10-min break to complete a similarity concept game (described below) and engage in a relaxing task (i.e., coloring). Next, the children repeated the task for the second suspect. Suspect order (male vs. female) was counterbalanced across participants, and photo presentation order was randomized to minimize any effects of fatigue. If children appeared tired, research assistants temporarily paused the session. Next, researchers debriefed the children and answered any questions. The full procedure took about 30 min to complete.

To choose the seven fillers for each lineup, we first used the adult pilot to compute a ‘gap’—the maximum minus the average similarity (e.g., 7.20–4.57 = 2.63). For each child target, we subtracted this gap from the child's maximum similarity to set a target similarity, then selected the seven faces whose similarity scores were closest to that target. Next, we applied the same dispersion pattern from the adult lineup to select fillers. For instance, if a filler in the adult lineup was +0.30 above the average, we added +0.30 to the child-derived average to select the corresponding filler. This process was repeated for all seven fillers. To match the structure of the adult-created lineup, the male innocent suspect was selected as the most similar filler, while the female innocent suspect was moderately similar. No fillers overlapped with those used in the adult-created lineups. All fillers were images available from the Glasgow Unfamiliar Face Database (Burton et al. 2010).

For each lineup, we computed (i) mean suspect–filler similarity, (ii) dispersion (range of suspect–filler similarity; SD when available), and (iii) innocent–suspect similarity (target-absent difficulty proxy). We summarize these metrics by experiment, target, and creator, and use them descriptively. Child-created arrays were less similar on average than adult-created and showed lower innocent–suspect similarity (male and female targets); dispersion was similar or slightly tighter for child-created (See Table 3). For the male target, child-created arrays had lower mean suspect–filler similarity than adult-created arrays (3.29 vs. 4.57), a slightly narrower range (2.14 vs. 2.47), and lower innocent–suspect (target-replacement) similarity (4.57 vs. 5.87). For the female target, the pattern was similar: lower mean similarity (2.76 vs. 3.24), comparable dispersion (2.71 vs. 2.67), and lower innocent–suspect similarity (2.95 vs. 3.32). See Table 3 for full similarity data and Figure 1 for the resulting lineup stimuli. We also calculated lineup fairness using the approach recommended by Mansour et al. (2017), and the lineup material was considered fair (0.92), such that the suspect did not stand out disproportionately in target-absent lineups. See Supporting Information for full details on lineup fairness analysis.

6.2.3 | Similarity Concept Game

To ensure the 22 children who provided similarity ratings understood the concept of similarity, they completed a brief concept game. Each child was shown four sets of items, each featuring a suspect image alongside four comparison options that varied in similarity. Children were provided with images of target items next to four images of a similar-target item: a jar of blue jelly beans (target) paired with four images of different colored jelly beans, a box full of soccer balls (target) paired with four images of boxes of different sport balls, a box full of bumblebees (target) paired with boxes full of different insects, and a plate full of strawberries (target) paired with plates full of different fruit. Each of the four paired items varied in the degree of similarity with the associated target item (e.g., each jar of colored jellybeans contained an increasing proportion of blue jellybeans). Children were tasked with selecting the target items that “looks most like” the target. The four tasks were presented in two different orders. All children but one completed the task with no guidance. The game was a brief proof-of-concept/engagement task to demonstrate that children in this age range could perform similarity judgments. It was not designed as an exclusion screen. Accordingly, we did not predefine exclusion criteria, and no exclusions were made based on this activity. We implemented the game only in Experiment 1 to establish feasibility.

6.2.4 | Suspect Videos

Two videos were shown—one of a male suspect and one of a female suspect, each 2 min in duration. The female video showed the female suspect enter a brightly lit room, pour a drink of water, sit and start reading. After some time passed, the female suspect spilled water on her notes and cleaned the mess before leaving the room. For the male suspect video, the suspect entered a room and began to study his notes. After a few minutes, he got up to do stretches and exercises before returning to studying. Each video contained a close-up of the suspect's face for approximately 20 s.

6.3 | Procedure

6.3.1 | Lineup Identification Procedure

A new sample of children ($N = 255$) participated in the identification task. Each week, on Day 1, groups of ~70 children (not all were participating) watched the videos of the male

TABLE 3 | Similarity rating differences in child-created and adult-created lineups.

		Male target		Female target	
		Child	Adult	Child	Adult
Experiment 1	Average similarity rating for all 100 faces	3.42	3.60	3.23	3.32
	Range 100 faces (min, max)	2.41	3.80	5.29	6.00
	Minimum 100 faces (min, max)	2.16	2.07	1.33	1.20
	Maximum 100 faces (min, max)	4.57	5.87	6.62	7.20
	Target replacement				
	Adult-created lineup	3.86	5.87	4.48	3.32
	Child-created lineup	4.57	3.60	2.95	4.53
	Range lineup				
	Adult-created lineup	2.41	2.47	2.14	2.67
	Child-created lineup	2.14	2.60	2.71	2.87
	Average lineup similarity rating				
	Adult-created lineup	3.28	4.57	4.69	3.24
	Child-created lineup	3.29	3.95	2.76	2.90
Experiment 2	Average similarity rating for all 100 faces	2.43	2.56	2.61	2.84
	Range 100 faces (max–min)	2.60	4.56	2.88	4.11
	Minimum 100 faces (min, max)	1.27	1.06	1.47	1.33
	Maximum 100 faces (min, max)	3.87	5.61	4.35	5.44
	Target replacement				
	Adult-created lineup	3.20	5.61	3.85	5.44
	Child-created lineup	3.87	4.17	4.35	4.67
	Range lineup				
	Adult-created lineup	1.23	1.94	0.95	1.72
	Child-created lineup	0.64	2.17	1.16	1.06
	Average lineup similarity rating				
	Adult-created lineup	2.40	2.55	2.73	2.78
	Child-created lineup	2.19	2.00	2.53	2.72
Experiment 3	Average similarity rating for all 50 faces	3.73	2.39	4.30	2.35
	Range 50 faces (min, max)	3.67	5.20	5.20	5.31
	Minimum 50 faces (min, max)	2.20	0.60	1.53	0.25
	Maximum 50 faces (min, max)	5.87	5.80	6.73	5.56
	Target replacement				
	Adult-created lineup	5.00	5.80	5.93	5.56
	Child-created lineup	5.87	3.53	6.73	2.44
	Range lineup				
	Adult-created lineup	2.33	4.06	2.86	3.69
	Child-created lineup	2.53	2.86	2.80	1.63
	Average lineup similarity rating				

(Continues)

TABLE 3 | (Continued)

	Male target		Female target	
	Child	**Adult**	**Child**	**Adult**
Adult-created lineup	3.59	2.58	4.51	2.77
Child-created lineup	4.58	2.86	4.60	2.32

and female suspects on a large auditorium projector, in a counterbalanced order. Children were instructed to watch carefully and were monitored during viewing. On Day 2, research assistants individually tested children with parental consent. After rapport-building, children completed two computer-administered simultaneous lineups, randomizing which lineup was presented first. Lineups were randomized to be target-absent or target-present, and the administering assistant was blind to target presence. Children received standardized, unbiased instructions (i.e., may or may not be present). After each identification and before any discussion, children rated their confidence on a scale of 1 to 11. The entire session took under 10 min per child. This Day 1 and Day 2 process was repeated over 5 weeks (Week 1, $n=66$; Week 2, $n=57$; Week 3, $n=59$; Week 4, $n=23$; Week 5, $n=56$). All sessions were audio recorded, and children received a small prize for participation.

7 | Experiment 1 Results

7.1 | Lineup Identification Decisions

Children made one of three identification responses: suspect identification, filler identification, or lineup rejection (see Table 4 for the complete pattern of responding).

Using this outcome information, we ran two analyses. First, responses were classified as correct or incorrect based on target presence: suspect identification (i.e., identifying the guilty suspect or culprit) was correct only in target-present lineups, lineup rejections were correct only in target-absent lineups, and filler selection was always incorrect. We then analyzed children's trial-level accuracy with a generalized linear mixed model (logit link) using glmer in lme4 (Baayen et al. 2008; Bates et al. 2015) in R (R Core Team 2020). The unreduced model predicted Correct (1) vs. Incorrect (0) from Age Group (6–8, 9–11), Lineup Creator (Child-created, Adult-created), and Target Presence (Present, Absent), including all interactions, with random intercepts for participant and actor/lineup. Accuracy was substantially lower in target-present than target-absent lineups ($b=-3.35$, SE = 0.60, $z=-5.59$, $p<0.001$; OR = 0.04, 95% CI [0.01, 0.11]). Two interactions qualified this effect. Presence × Age: older children (9–11) outperformed younger children (6–8) in target-present arrays, with little difference in target-absent arrays ($b=1.62$, SE = 0.72, $z=2.26$, $p=0.024$; OR = 5.05, 95% CI [1.23, 20.49]). Presence × Creator: child-created lineups yielded higher accuracy in target-present arrays than adult-created lineups, with no difference in target-absent arrays ($b=1.49$, SE = 0.75, $z=1.98$, $p=0.047$; OR = 4.44, 95% CI [1.02, 19.30]). Main effects of Creator ($b=-0.35$, $p=0.386$) and Age ($b=-0.68$, $p=0.066$), the Creator × Age interaction ($b=0.96$, $p=0.071$), and the three-way interaction ($b=-1.18$, $p=0.201$) were not reliable once interactions with Presence were modeled. Random-effects variances indicated negligible between-participant intercept variance (SD = 0.0001) and meaningful lineup/actor variance (SD = 0.34). The model used 508 observations from 255 participants across 2 actor/lineup levels.

Next, we also fit a multinomial logistic regression predicting decision type (suspect, filler, rejection) from Target Presence (TP vs. TA), Creator (Adult vs. Child), and their interaction, with participant-cluster bootstrap 95% CIs for inference. Decisions varied with target presence, and the Creator × Presence pattern indicated greater selectivity for child-created lineups (fewer suspect IDs when the target was absent and more suspect IDs when the target was present) relative to adult-created lineups. See Table 5 for predicted probabilities for suspect identifications in each cell for all experiments. Tables S9 and S10 provide the full regression results, including the odds ratios and complementary filler/reject predicted probabilities.

7.2 | Discriminability Information

We also completed signal-detection (d') for suspect identifications within each Lineup-creator condition (Adult- vs. Child-created), treating a suspect identification as a “yes” response (Mickes et al. 2014). The child-created condition showed clear separation of TP and TA decisions, whereas the adult-created condition showed little to no discriminability in this sample. See Table 6 for d' estimates and bootstrap 95% CIs.

7.3 | Confidence

We analyzed trial-level confidence ratings to estimate confidence–accuracy characteristic (CAC) functions for suspect identifications and to compute the positive predictive value (PPV) at each confidence level, separately for adult-created and child-created lineups (see Figure 2). Confidence (1–11 scale) was binned into Low (1–3), Medium (4–7), and High (8–11) for primary figures; for each Creator × confidence bin, we computed PPV and Wilson 95% CIs based on the proportion of correct suspect IDs. Bins with too few suspect IDs (pre-specified threshold) were not plotted to avoid unstable estimates. As seen in Figure 2, at Medium confidence (4–7), PPV was 57.9% [36.3, 76.9] for Adult-created ($n=19$) versus 87.0% [67.9, 95.5] for Child-created ($n=23$). At High confidence (8–11), PPV was 50.0% [23.7, 76.3] for Adult-created ($n=10$) versus 87.5% [64.0, 96.9]

TABLE 4 | Lineup identification decisions for all experiments.

					Identification procedure			
Exp.	**Age group**	**Stimuli**	**Suspect**	**Presence**	**Suspect**	**Filler**	**Reject**	***n***
Exp. 1	Child	Adult-Created[a]	Male	Present	0.13	0.36	0.52	62
				Absent	0.16	0.37	0.46	67
			Female	Present	0.13	0.33	0.54	63
				Absent	0.06	0.22	0.72	67
		Child-created	Male	Present	0.18	0.18	0.65	61
				Absent	0.05	0.37	0.59	63
			Female	Present	0.38	0.22	0.41	64
				Absent	0.03	0.28	0.69	61
Exp. 2	Child	Adult-created	Male	Present	0.62	0.10	0.029	52
				Absent	0.17	0.17	0.67	72
			Female	Present	0.50	0.11	0.39	64
				Absent	0.14	0.28	0.58	64
		Child-created	Male	Present	0.59	0.15	0.25	59
				Absent	0.05	0.20	0.75	55
			Female	Present	0.57	0.12	0.31	67
				Absent	0.08	0.11	0.81	53
Exp. 3	Adult	Adult-created	Male	Present	0.54	0.14	0.32	37
				Absent	0.07	0.30	0.63	30
			Female	Present	0.63	0.16	0.21	38
				Absent	0.03	0.25	0.72	36
		Child-created	Male	Present	0.60	0.07	0.33	45
				Absent	0.07	0.10	0.83	30
			Female	Present	0.47	0.28	0.25	32
				Absent	0.03	0.30	0.67	36
Exp. 3	Child	Adult-created	Male	Present	0.13	0.41	0.46	37
				Absent	0.07	0.33	0.60	30
			Female	Present	0.24	0.16	0.60	37
				Absent	0.13	0.20	0.67	30
		Child-created	Male	Present	0.15	0.39	0.67	46
				Absent	0.19	0.39	0.42	31
			Female	Present	0.17	0.28	0.54	46
				Absent	0.11	0.26	0.63	27

[a] Unlike the adult-created lineup data reported in Table 2, these data are not previously reported.

for Child-created ($n = 16$). Collapsed across confidence, overall PPV was 51.6% [34.8, 68.0] for Adult-created ($n = 31$) versus 87.5% [73.9, 94.5] for Child-created ($n = 40$). Overall, Experiment 1 shows higher PPV for Child-created lineups at medium/high confidence. Full confidence analysis results can be found in Table S5 for all experiments.

8 | Experiment 1 Discussion

We hypothesized that building lineups using children's perceptions of suspect-filler similarity would improve performance relative to adult-created lineups. The results generally support this hypothesis. Children shown a child-created lineup produced a pattern of responding consistent with more diagnostic responding in target-present lineups—that is, more suspect identifications were made when the target was present and fewer when absent. Discriminability (d′) was higher for child-created than adult-created lineups, indicating better separation of target-present and target-absent decisions. Confidence–accuracy curves also suggest that child-created lineups resulted in confidence ratings that were more reflective of accuracy responding, particularly at higher levels of confidence.

TABLE 5 | Predicted probability of a suspect identification by target presence (TP/TA) and lineup creator (adult- vs. child-created) for each experiment.

Experiment	Presence	Creator	*P* (choosing the suspect) with 95% CI
Experiment 1	TA	Adult-created	0.11 [0.06, 0.18]
	TA	Child-created	0.04 [0.01, 0.08]
	TP	Adult-created	0.13 [0.08, 0.19]
	TP	Child-created	0.28 [0.20, 0.36]
Experiment 2	TA	Adult-created	0.15 [0.09, 0.23]
	TA	Child-created	0.06 [0.02, 0.12]
	TP	Adult-created	0.55 [0.44, 0.66]
	TP	Child-created	0.58 [0.47, 0.68]
Experiment 3—adult	TA	Adult-created	0.05 [0.00, 0.10]
	TA	Child-created	0.05 [0.00, 0.10]
	TP	Adult-created	0.59 [0.47, 0.70]
	TP	Child-created	0.55 [0.44, 0.66]
Experiment 3—child	TA	Adult-created	0.10 [0.03, 0.18]
	TA	Child-created	0.16 [0.07, 0.25]
	TP	Adult-created	0.19 [0.11, 0.29]
	TP	Child-created	0.16 [0.09, 0.24]

Note: Predicted probabilities are from a multinomial logistic regression of decision type (suspect, filler, reject) on Target Presence (TP vs. TA), Lineup Creator (child- vs. adult-created), and their interaction, fit with "nnet::multinom" in R. The baseline outcome is reject; predictions are reported on the probability scale. Inference uses a participant-cluster bootstrap ($B = 2000$ resamples), resampling participants with replacement and refitting the model each time. 95% CIs are percentile bootstrap intervals.

However, it is notable that correct suspect identification rates in target-present lineups were lower than typically observed in similar age groups (e.g., Fitzgerald et al. 2014). This unexpected pattern may suggest that task difficulty or related attentional factors—perhaps using a video as the encoding stimulus—contributed to children's performance. Viewing the stimuli video in large groups with less engaging material (video) likely created an unideal encoding environment. As such, we conducted Experiment 2 to replicate these findings with a more naturalistic encoding design.

9 | Experiment 2

Experiment 2 examined whether we would replicate the results of Experiment 1 when using: (a) child- and adult-created lineup stimuli developed independently from each other (different from Experiment 1 method) and (b) enriched encoding conditions that may be more reflective of real-world scenarios—that is, using live suspects with increased exposure time. The latter was important to explore because enriched encoding conditions may give children the support needed to overcome the observed differences across child- and adult-created lineups observed in Experiment 1. With richer encoding conditions, we anticipated overall higher accuracy. Older children were expected to outperform younger ones, but child-created lineups were predicted to reduce younger children's false identifications and narrow the age gap.

10 | Experiment 2 Method

10.1 | Participants and Design

To create the study materials, we recruited 20 children (7- to 12-years-old; $M_{\text{age}} = 9.70$, $\text{SD} = 1.75$; 50% female) and 18 adults from the community ($M_{\text{age}} = 27.39$, $\text{SD} = 12.22$; 83% female) to provide similarity ratings. For the main study, we recruited 249 new children, 6- to 11-years-old ($M_{\text{age}} = 8.63$, $\text{SD} = 1.42$; 45%

TABLE 6 | Discriminability (d′) across experiments.

Experiment	Creator	TP *n*	TA *n*	TP hits	TA suspect IDs	d′ [95% CI]	H (TP)	FA (TA)
Experiment 1	Adult	125	134	16	15	0.08 [−0.32, 0.47]	0.13	0.11
	Child	125	124	35	5	1.13 [0.72, 1.68]	0.28	0.04
Experiment 2	Adult	116	136	64	21	1.14 [0.80, 1.49]	0.55	0.16
	Child	126	108	73	7	1.68 [1.29, 2.19]	0.58	0.07
Experiment 3—adult	Adult	75	66	44	3	1.84 [1.35, 2.65]	0.59	0.05
	Child	77	66	42	3	1.74 [1.26, 2.51]	0.54	0.05
Experiment 3—child	Adult	74	60	14	6	0.38 [−0.15, 0.97]	0.19	0.11
	Child	92	58	15	9	0.02 [−0.42, 0.53]	0.17	0.16

female) from a science camp. The SES and ethnic makeup of our sample were like those in Experiment 1. Children were assigned to a 2 (Lineup-creator: Child, Adult) × 2 (Actor: Male, Female) × 2 (Age: 6–8, 9–11) × 2 (Target Presence: Present, Absent) mixed design in which Target Presence of each suspect was randomly assigned by computer for each participant. Post hoc sensitivity (80% power, $\alpha = 0.05$; See Supporting Information) showed the minimum detectable Target Presence × Lineup-Creator interaction on accuracy was 1.39 log-odds (OR = 4.03; $n = 486$, PP = 215). For the three-way interaction with Age Group, the MDES was 2.88 log-odds (OR = 17.80), again reflecting low sensitivity to higher-order effects.

10.2 | Materials

10.2.1 | Lineup Stimuli

We developed our lineup stimuli for Experiment 2 using a standardized procedure. Child and adult judges rated the similarity between each suspect (1 male, 1 female) and 100 potential fillers (matched on gender and race; Glasgow Unfamiliar Face Database; Burton et al. 2010) using a 10-point Likert-type scale (0 = not at all similar, 10 = highly similar). For both suspects, the innocent suspect was selected as the most similar filler. This reflects a high-risk, real-world scenario where innocent suspects resemble the culprit closely (Navon 1992; Wells and Penrod 2011). This approach also aligns with past developmental eyewitness studies, enhancing comparability (e.g., Bruer et al. 2017; Fitzgerald et al. 2014; Karageorge and Zajac 2011; Price et al. 2020). To select the remaining seven fillers, we randomly chose faces within ±0.5 SD of the overall similarity rating mean, ensuring moderate similarity (see Fitzgerald et al. 2013). The *child-created lineup* stimuli were selected based on children's similarity ratings, while the *adult-created lineup* stimuli were selected using adult similarity ratings. No fillers overlapped between the child- and adult-created lineup stimuli.

Full details on the similarity rates associated with the lineup stimuli can be found in Table 3 and the resulting lineup stimuli can be seen in Figure 3. Child-created arrays again showed lower mean similarity, tighter dispersion, and lower innocent–suspect similarity than adult-created (both targets; Table 3). For the male target, child-created arrays showed lower mean similarity (2.19 vs. 2.55), a much tighter range (0.64 vs. 1.94), and lower innocent–suspect similarity (3.87 vs. 5.61). For the female target, child-created arrays likewise had lower mean similarity (2.53 vs. 2.78), a tighter range (1.16 vs. 1.72), and lower innocent–suspect similarity (4.35 vs.

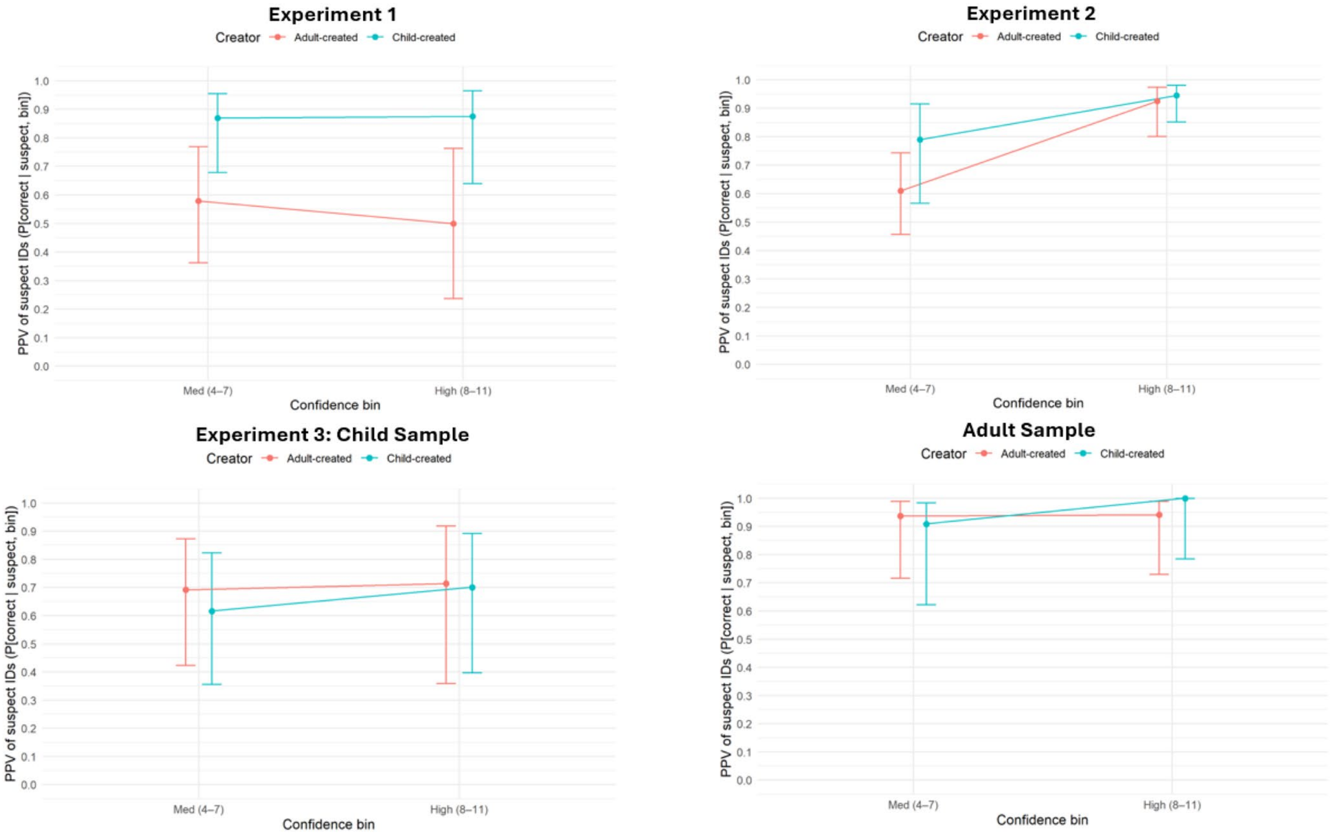

FIGURE 2 | Confidence–accuracy calibration (CAC) for suspect identifications by lineup creator and confidence. *Note:* Points are PPV = P(correct|suspect ID) within confidence bins; error bars are 95% Wilson CIs. Confidence was collected on a 1–11 scale and binned as Low (1–3), Medium (4–7), and High (8–11). The Low-confidence bin is absent in situations when virtually no suspect identifications were made with ratings of 1–3 and/or the few that occurred did not meet the pre-specified stability rule (bins are plotted only when each creator has ≥ 10 suspect IDs in that bin). Experiment 3 bins used relaxed-threshold versions of these panels (bins shown when a creator has ≥ 3 suspect IDs) to have enough in the bins to visualize in a comparable way with other experiments. Trials with missing confidence were excluded from binning.

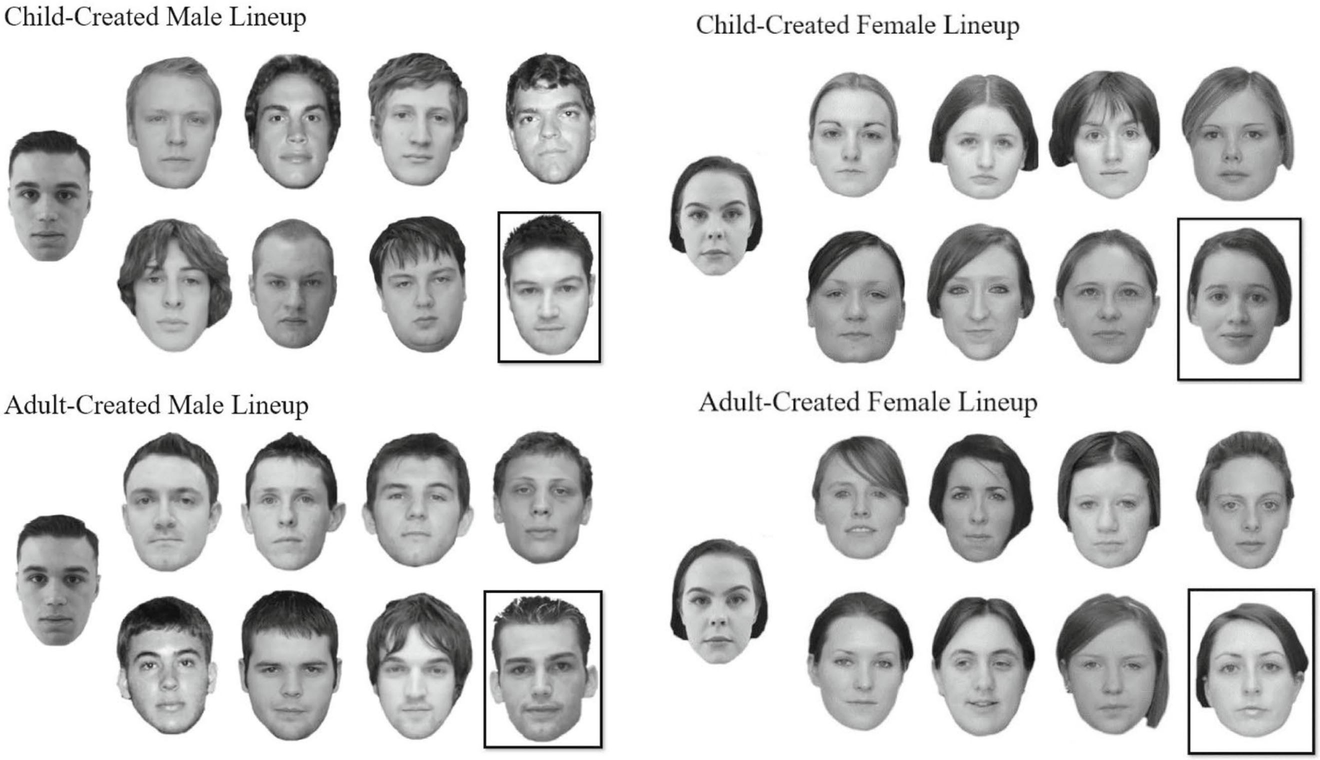
Child-Created Male Lineup
Child-Created Female Lineup
Adult-Created Male Lineup
Adult-Created Female Lineup

FIGURE 3 | Experiment 2 lineup stimuli. *Note:* Target is on the left and the target replacement is indicated with the black box.

5.44). All lineup images were displayed on an 11-in. touchscreen tablet and shown using *OpenSesame* software that recorded participants' touch responses. As done in Experiment 1, lineup fairness was assessed, and these lineups were generally fair. See Supporting Information: Section 1.0 for lineup fairness information.

10.3 | Procedure

Children participated in a live event in which a female and male research assistant (the suspects, both Caucasian in their early 20s) visited children in groups of approximately 10–15 children. After introducing themselves as artists, the suspects conducted a science-based art show. After the first trick, the visitors 'accidentally' spilled water on a computer, resulting in the computer breaking and them ending the demonstration early. The interactions were audio recorded to ensure consistency across groups, and each session lasted about 10 min. The following day, different research assistants worked with children individually. After rapport building and receiving children's assent, children completed the lineup tasks (one for the male suspect and one for the female suspect). The same lineup administration procedure described in Experiment 1 was used (i.e., unbiased instructions, counterbalanced lineups, simultaneous presentation) and interviews were audio recorded.

11 | Experiment 2 Results

11.1 | Lineup Identification Decisions

See Table 3 for a complete breakdown of lineup identification decisions across conditions (collapsed across age). We first ran the same GLMM as described in Experiment 1 (see full results in Tables in Supporting Information). The GLMM yielded a significant main effect of Lineup Creator: children were more accurate with child-created lineups than with adult-created lineups ($b = 1.11$, SE = 0.55, $z = 2.00$, $p = 0.045$; OR = 3.03, 95% CI [1.02, 8.93]). No other fixed effects were reliable: Target Presence ($b = -0.36$, SE = 0.45, $z = -0.79$, $p = 0.430$; OR = 0.70, 95% CI [0.29, 1.70]) and Age Group ($b = 0.23$, SE = 0.51, $z = 0.46$, $p = 0.649$; OR = 1.26, 95% CI [0.47, 3.39]) were nonsignificant, and interactions with Presence and Age did not reach significance (largest $|z| = 1.59$, all $ps \geq 0.113$). Random-effects estimates indicated substantial between-participant variability (SD = 1.24) and negligible actor/lineup variance (SD = 0.00001). The model included 486 observations from 215 participants across 2 actor/lineup levels.

Next, we ran the same multinomial model described in Experiment 1. As seen in Experiment 1, target presence strongly influenced decisions. Relative to adult-created lineups, child-created lineups showed reduced suspect identifications in target-absent arrays and comparable or slightly higher suspect identifications in target-present arrays. See Table 5 for predicted probabilities for suspect identifications in each cell for all experiments. Tables S9 and S10 provide the full regression results, including the odds ratios and complementary filler/reject predicted probabilities.

11.2 | Discriminability Information

Both lineup types produced above-zero discriminability, with child-created lineups showing a consistent d' advantage relative to adult-created lineups. Although intervals overlap, the pattern favors child-created stimuli while confirming that both lineup types enabled meaningful TP–TA separation. See Table 6 for exact d′ values and 95% CIs.

11.3 | Confidence

We conducted the same analysis described in Experiment 1 to PPV at each confidence level, separately for adult-created and child-created lineups. As seen in Figure 1, PPV curves show the expected pattern that PPV for suspect IDs increased as confidence increased, and high-confidence suspect IDs were most accurate. PPV curves for adult-created and child-created lineups were broadly similar, with overlapping CIs.

12 | Experiment 2 Discussion

In Experiment 2, we examined whether we could replicate the benefits observed in Experiment 1 with different suspects using a design more conducive to a richer encoding experience (i.e., live event, smaller group, longer exposure) that may be more reflective of real-world eyewitness experiences. The results generally converge on the same conclusion: child-created lineups yielded greater discriminability than adult-created lineups. The results indicated a pattern of responding that suggests improved discriminability—that is, a higher propensity to choose the suspect when the target was present than when the target was absent. However, in the experiments conducted so far, no adult sample was used. Including adults allows us to disentangle child-specific effects from general stimulus properties, providing critical insight into whether lineup design should vary by age to optimize eyewitness performance.

13 | Experiment 3

The purpose of Experiment 3 was twofold. First, we aimed to replicate the benefits of child-created lineups observed in Experiments 1 and 2 using a new set of suspect faces. Second, we included an adult comparison group to examine whether the benefits of child-created lineups extend to adult witnesses. If adults also show improved accuracy with child-created lineups, this may signal the need to take a closer look at what a child-created lineup comprises to see if we can replicate those characteristics in future lineups. Conversely, if child-created lineups offer no benefit—or even impair performance—for adults, this would suggest the effects are specific to children's perceptual or cognitive processing, strengthening the argument that lineup construction should be developmentally tailored. Given that we found that child-created stimuli had benefits for children who were exposed to suspect faces using both video (Experiment 1) and live (Experiment 2) stimuli, for Experiment 3, we opted to use video stimuli to ensure comparable encoding conditions across the child and adult samples. For children, we predicted that older children would outperform younger ones, with younger children again showing more benefit from child-created lineups. Adults were expected to perform well regardless of lineup type, with age-group differences most evident in target-present conditions.

14 | Experiment 3 Method

14.1 | Participants and Design

To create the study materials, we recruited 15 children (6- to 12-years-old; $M_{age} = 9.74$, SD = 2.10; 40% female) and 16 adults (18- to 58-years-old; $M_{age} = 31.00$, SD = 15.69; 81% female) from the community to provide similarity ratings. We then recruited 144 new children, 8- to 12-years-old ($M_{age} = 9.86$, SD = 1.42; 45% female) from a camp, recruited from a community with similar SES and ethnicity to those described in the previous experiments. These children were assigned to a 2 (Lineup-creator: Child, Adult) × 2 (Actor: Male, Female) × 2 (Age: 8–9, 10–12) × 2 (Target Presence: Present, Absent) mixed design in which the presence of each suspect was randomly assigned by a computer for each participant. Additionally, we recruited 142 adults ($M_{age} = 22.65$, SD = 5.93; 63% female) from an undergraduate research participant pool. Ethnicity was most reported as Caucasian (47%), South Asian (24%), Southeast Asian (9%), or East Asian (7%). Undergraduate participants were assigned to a 2 (Lineup-creator: Child, Adult) × 2 (Actor: Male, Female) × 2 (Target Presence: Present, Absent) mixed design in which the presence of each suspect was randomly assigned for each participant. For the child sample, post hoc sensitivity (80% power, $\alpha = 0.05$) indicated the minimum detectable Target Presence × Lineup-Creator interaction on accuracy was 1.62 log-odds (OR = 5.03; $n = 284$, PP = 142). For the adult sample, the minimum detectable Target Presence × Lineup-Creator interaction on accuracy was 1.42 log-odds (OR = 4.13; $n = 284$, $PP = 141$).

14.2 | Materials

14.2.1 | Suspect Video

Like Experiment 1, small groups of participants watched one video containing two suspects, a male and a female, entering a room. After exploring the room, the suspects found some vision impairment goggles, put on the goggles, and tried tossing a ball back and forth. The video length was 1 min and 52 s. The two suspects were different people but of similar age to those used in Experiment 1 and Experiment 2 (i.e., both suspects were Caucasian in their early 20s).

14.2.2 | Lineup Stimuli

A similar method was used to create lineup stimuli as was described in Experiment 2—with one notable difference: child ($n = 15$) and adult judges ($n = 16$) provided pairwise (subjective) similarity ratings of photographs of the two suspects, each with only 50 potential fillers. As seen in Table 3, in contrast to Experiments 1 and 2, child-created lineups were more similar (and less dispersed) than adult-created, with equal or higher innocent–suspect similarity. For the male target, child-created arrays had higher mean similarity than adult-created (4.58 vs. 2.58), a narrower range (2.53 vs. 4.06), and slightly higher innocent–suspect similarity (5.87 vs. 5.80). For the female target, child-created arrays also had higher mean similarity (4.60 vs. 2.77), a narrower range (2.80 vs. 3.69), and higher

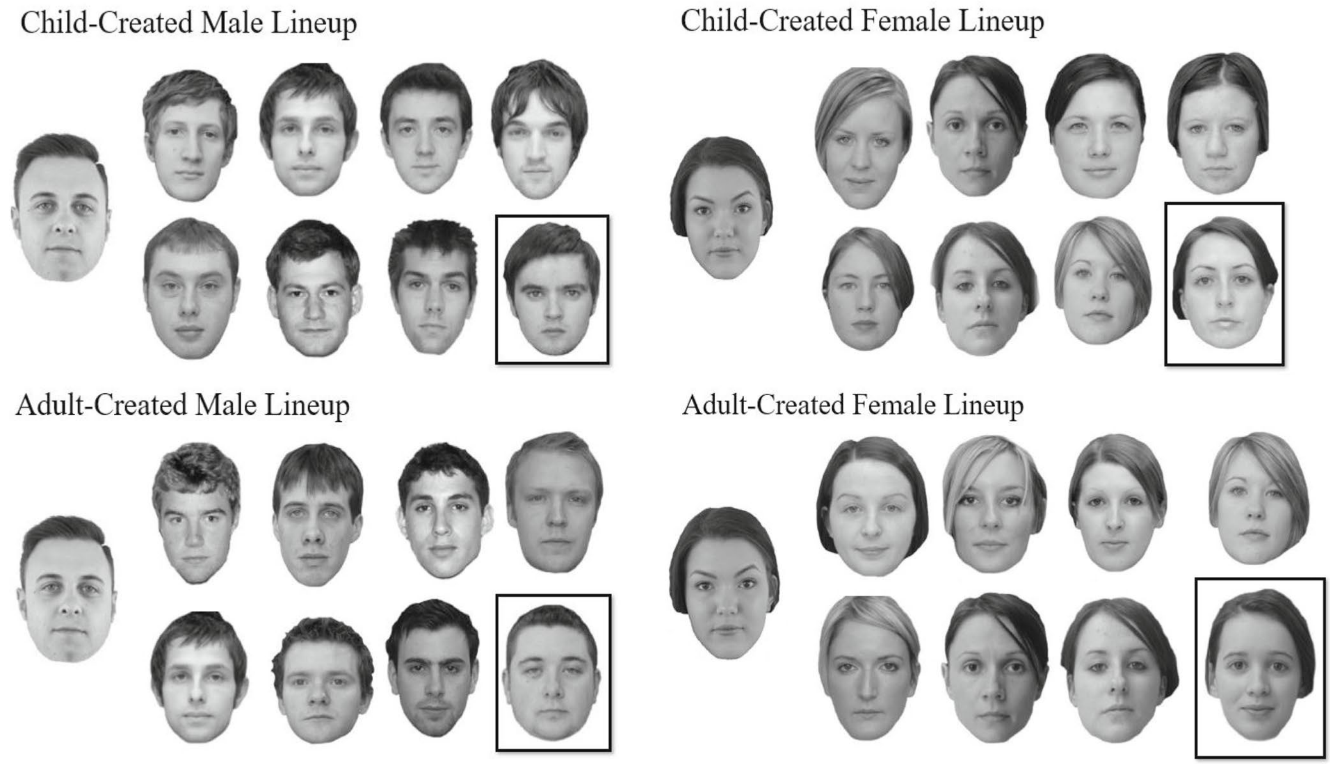
Child-Created Male Lineup
Child-Created Female Lineup
Adult-Created Male Lineup
Adult-Created Female Lineup

FIGURE 4 | Experiment 3 lineup stimuli. *Note:* Target is on the left and target replacement is indicated with the black box.

innocent–suspect similarity (6.73 vs. 5.56). See Supporting Information for information on lineup fairness. Two fillers overlapped between the child- and adult-created lineup stimuli, which is shown in Figure 4.

14.3 | Procedure

14.3.1 | Child Procedure

On the first day, children watched the above-described video. The following day, children were individually paired with a research assistant. The same lineup administration procedure described in Experiments 1 and 2 was used (i.e., unbiased instructions, counterbalanced lineups, simultaneous presentation). Participants viewed the video of the male and female suspect and then completed two identification tasks, one for each suspect. The order of lineups was counterbalanced. The interview with the child was audio-recorded. All lineup images were displayed on an 11-in. touch screen tablet and shown using *Eprime 2.0* software that recorded participants' responses.

14.3.2 | Adult Procedure

Participants completed the lineup task electronically (via Qualtrics), either in the lab ($n = 52$) or, due to social distancing restrictions present during the pandemic, online ($n = 90$). After reviewing the consent form, adults watched the suspect video. Participants were told to watch the video closely as they would be asked to make judgments about the actors' personalities. After providing basic demographic information, participants received a second consent form that indicated that the study was about eyewitness memory. After reconfirming their consent, they were provided with two lineups (one for the male and female suspects). The order of the lineups was counterbalanced. Target presence and location of the suspect in the lineup were randomly determined by the computer. Participants were also provided with written, unbiased instructions, indicating that the people from the video may or may not be present. Once each lineup was completed, participants who completed the procedure online provided a confidence rating on a scale of 1 to 11 ($n = 90$). Participants received course credit for their participation.

15 | Experiment 3 Results

15.1 | Lineup Identification Decisions

Children: We ran the same GLMM as described in Experiment 1. Inferential statistics of the model are presented in Table S8A. The only reliable effect was Target Presence: children were less accurate in target-present than target-absent arrays, replicating the Experiment 1 pattern with video stimuli ($b = -2.22$, $SE = 0.58$, $z = -3.83$, $p < 0.001$; $OR = 0.11$, 95% CI [0.03, 0.34]). Main effects of Creator and Age and all interactions (including Presence × Creator and Presence × Age) were not significant. Random-effects estimates indicated between-participant variability ($SD = 0.57$) and actor/lineup variability ($SD = 0.15$). The model included 284 observations from 142 participants across 2 actor/lineup levels.

We also ran the same multinomial regression described in Experiment 1 and found that, in the child sample, child-created lineups tended to yield slightly more suspect IDs in target-absent arrays and slightly fewer in target-present arrays than adult-created lineups. Full results can be seen in Table 5.

Adults: A similar GLMM was run (without the Age variable, as this sample did not have different age groups) (Table S8B). Neither Target Presence ($b = -0.41$, $SE = 0.35$, $z = -1.17$, $p = 0.244$; $OR = 0.66$, 95% CI [0.33, 1.32]) nor Lineup Creator ($b = 0.30$, $SE = 0.39$, $z = 0.77$, $p = 0.443$) nor their interaction ($b = -0.46$, $SE = 0.51$, $z = -0.92$, $p = 0.359$) reached significance. Random-intercept variances were effectively zero (boundary fit), consistent with minimal between-cluster variability in this dataset. The model included 284 observations from 141 participants across 2 actor/lineup levels. In the second analysis (see Table 5), the adult sample's decision distributions were very similar across creators.

15.2 | Discriminability Information

Children: Among child witnesses, discriminability was low and imprecisely estimated for both lineup types; bootstrap intervals include zero, indicating limited TP–TA separation regardless of creator. See Table 6 for estimates and 95% CIs.

Adults: Among adult witnesses, both lineup types yielded high discriminability, with no clear difference between child-created and adult-created lineups. See Table 6 for estimates and 95% CIs.

15.3 | Confidence

Children: The same confidence previously described was run and, as seen in Figure 2, PPV did not differ between adult- and child-created stimuli; however, the PPV was lower overall than that of adults.

Adults: As seen in Figure 2, PPV also did not differ according to lineup creator.

16 | Experiment 3 Discussion

Different from Experiments 1 and 2, we did not observe an advantage of using child-created lineups with child eyewitnesses in Experiment 3. Discriminability estimates (d') were low and imprecise for both lineup types in the child sample, and we observed a similar or less favorable decision and accuracy pattern for child-created lineups (i.e., slightly more suspect IDs in TA arrays and slightly fewer in TP arrays) relative to adult-created lineups; CAC curves showed the same direction, with wide intervals. For the adult sample, the similarity ratings used when selecting fillers did not influence identification responses, nor the discriminability of those identifications. We address this discrepancy more fully in the General Discussion.

17 | Meta-Analytic Integration Across Experiments

Given that all three experiments examined the influence of lineup construction method (child-created vs. adult-created) on children's identification accuracy, we conducted an exploratory meta-analytic integration to estimate the overall effect size across studies. A random-effects meta-analysis using the *meta* package in R yielded a pooled odds ratio of 1.37, 95% CI [0.71, 2.64], $z = 0.94$, $p = 0.35$, suggesting a small but non-significant advantage for child-created lineups in improving suspect identification accuracy. As seen in Figure 5, while Experiment 1 showed a significant benefit for child-created lineups, the others produced null or inconsistent findings.

Heterogeneity across studies was moderate ($I^2 = 66.5\%$, $Q(2) = 5.97$, $p = 0.050$), indicating variability in the size and direction of effects. To better understand this variability in effects across studies, we examined potential sources of heterogeneity. A leave-one-out sensitivity analysis revealed that Experiment 1 substantially influenced the pooled effect. When it was excluded, the overall effect size dropped to near zero and was non-significant ($OR = 1.03$, 95% CI [0.67, 1.57], $p = 0.90$), and residual heterogeneity was eliminated ($I^2 = 0\%$). In contrast, removing Experiment 2 or 3 had minimal impact on the overall estimate or heterogeneity, suggesting that the main source of variability stemmed from Experiment 1. To explore possible moderators, we conducted exploratory meta-regressions to test whether participant age or encoding medium (live vs. video) accounted for differences in effect sizes. Neither age ($\beta = -0.58$, $SE = 0.71$, $z = -0.81$, $p = 0.42$) nor encoding medium (i.e., video vs. live; $\beta = 0.30$, $SE = 0.95$, $z = 0.32$, $p = 0.75$) significantly predicted the effect of lineup construction method. Both models left substantial residual heterogeneity unexplained ($I^2 > 75\%$), indicating that additional factors may contribute to variability across studies. These moderator findings should be interpreted cautiously,

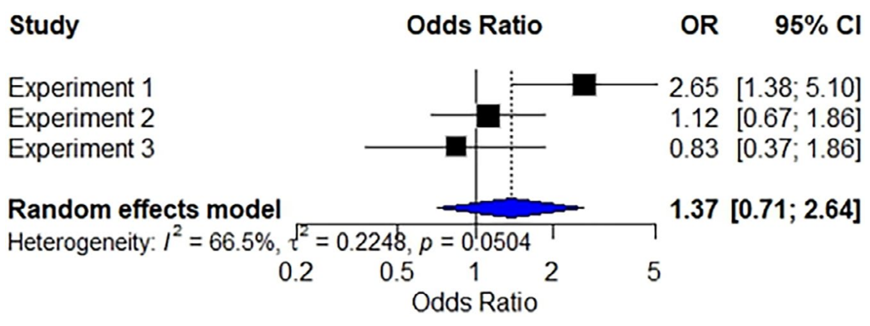

FIGURE 5 | Meta-analysis results.

given the small number of studies and the exploratory nature of the analysis.

17.1 | Mechanism Exploration

Because lineup creator and lineup similarity could not be included in the same models without producing unstable estimates, we conducted follow-up analyses at the lineup level, pooling across experiments. We asked whether the observed creator differences were a function of variability in suspect-filler similarity. Lineups with higher mean suspect-filler similarity resulted in less accurate identification decisions: a logistic regression showed lower accuracy for high versus low similarity lineups, $\chi^2(1) = 29.04$, $p < 0.001$, and the overall association (Spearman's correlation) was strong ($\rho = -0.58$, $p = 0.005$). In contrast, innocent suspect-filler similarity was not consistently related to accuracy; the overall association was negligible ($\rho = -0.03$, $p = 0.889$), although a median-split test suggested a small and opposite pattern, with slightly higher accuracy for high-similarity lineups (difference = +0.08), $\chi^2(1) = 4.69$, $p = 0.030$. These results indicate that the apparent creator effect is best understood as a function of lineup similarity: child- and adult-created lineups differed systematically in how similar the faces were, and this similarity, in turn, shaped identification accuracy.

Of course, it is important to note that because child-created lineups were built from child ratings and adult-created lineups from adult ratings, part of the difference could be due to how children versus adults use the rating scale. To address this potential circularity, we re-expressed similarity on a within-rater-group z metric (standardized or comparable scale) and examined cross-rater convergence when adults and children rated the same stimuli. When adults and children rated the same lineups, their ratings showed little agreement for mean suspect–filler similarity (Spearman's $\rho = 0.11$), moderate agreement for dispersion (range; $\rho = 0.58$), and systematic disagreement for innocent–suspect similarity ($\rho = -0.46$). In other words, where adults perceived higher innocent–suspect similarity, children tended to perceive lower similarity, and vice versa. Therefore, the "creator" differences are best explained by lineup similarity in these materials; however, a definitive test will require similarity ratings from independent, age-neutral raters of the final arrays.

18 | General Discussion

Does considering children's perceptions of similarity when building a lineup impact their identification responding and how diagnostic their responding is of suspect guilt/innocence? It might. An examination of the current developmental lineup research revealed that most researchers used adult-provided similarity ratings when building lineups for use with child eyewitnesses. The present research examined whether incorporating a child's perceptions of similarity when building a lineup influences children's accuracy on lineup tasks.

In our pilot research, we present evidence that children rated similarity between pairs of faces differently from adults. Experiment 1 was designed to experimentally explore whether including children's perceived similarity in the design of lineup stimuli influenced child eyewitnesses' accuracy on a lineup following a brief video encoding experience. Experiment 2 expanded our exploration to test the influence of using child-created lineups with children who were exposed to two new suspect faces with a strong encoding context (i.e., a live event, 10-min exposure). Lastly, Experiment 3 explored how using child-created (versus adult-created stimuli) influenced both child and adult eyewitnesses. Below, we outline four key insights that emerged from this body of work.

First, children as young as 6 to 12 years old were able to rate facial similarity in ways that reliably influenced lineup construction and decision-making. Their ratings, while different from adults', produced lineups that, in some cases, yielded a more diagnostic pattern of responses—especially for the youngest children in our samples. This suggests children aged 6–11 generally possess metacognitive insight into perceptual similarity—a concept that has traditionally been assumed to require more mature cognitive development (Brown and Kane 1988).

Second, children perceive and subsequently rate similarity between faces differently than adults (see Table 3). In Experiment 3, children consistently assigned higher similarity ratings to face stimuli; therefore, their innocent suspects were perceived as more similar to the guilty suspect than those selected by adults. Across Experiments 2 and 3, child-created lineups were more perceptually clustered, with tighter ranges of similarity scores. These patterns suggest that children tend to construct more homogeneous lineups that may increase identification difficulty, particularly when memory strength is low (e.g., poor encoding conditions). There is ample research to support the notion that configural processing of face information in recognition (i.e., considering the spacing between facial features) develops throughout childhood and into the teen years (e.g., Mondloch et al. 2003). Children and adults both draw on featural and multiple forms of configural cues. What changes developmentally is the relative weighting of these routes, which become more adult-like gradually and depend on task demands and stimuli (e.g., if the face is unfamiliar or not). Younger children often weigh featural/external cues (e.g., hair) more when recognizing unfamiliar adult faces, whereas older children rely more on internal features (eyebrows, eyes, nose, mouth, and inner cheek; Ge et al. 2008; Want et al. 2003; but see Wilson et al. 2007 for recognition of familiar faces). Perhaps when rating similarity between pairs of faces, children in the present research relied more on exterior features (e.g., hair) while adults may have relied on a more configural or holistic approach to rating faces.

Third, in two of the three experiments (Experiments 1 and 2), we observed benefits of a lineup that was developed using similarity ratings provided by children (rather than adults) with child eyewitnesses. Importantly, we found evidence of benefits across multiple suspects/stimuli, using both video and live-suspect exposure methods. In Experiment 1, using a video-suspect exposure method, children shown child-created lineups were more likely to identify the suspect when the target was present, without a corresponding increase in innocent suspect identifications when the target was absent—a pattern consistent with better discriminability. Experiment 2 replicated this advantage using a live-suspect exposure method, again showing higher suspect identifications in target-present lineups and comparable or lower suspect identifications in target-absent lineups for child-created lineups.

Of note, across studies, the confidence-accuracy relationship for children's suspect identifications showed a general pattern whereby higher confidence corresponded to a higher probability that a suspect identification was correct. In Experiments 1 and 2, CAC results indicated that child-created lineups yielded more diagnostic high-confidence suspect identifications than adult-created lineups, with no corresponding rise in high-confidence errors in target-absent arrays. Thus, the creator effect discussed earlier is best characterized as an improvement in diagnostic responding at higher confidence, not a general liberalization of choosing. These CAC patterns converge with the d' and multinomial results, which likewise point to better separation of target-present and target-absent decisions for child-created lineups.

Our findings also suggest that the apparent advantage for child-created lineups observed in the two experiments is best explained by associated differences in lineup similarity. Greater suspect–filler similarity consistently predicted lower accuracy, consistent with prior work highlighting lineup fairness as a critical determinant of identification outcomes (Clark 2012; Wells et al. 2020). Because similarity was confounded with creator in the current stimuli, future research should vary creator and similarity independently to more clearly establish their separate contributions.

In Experiment 3, however, we did not observe an advantage of child-created lineups for children, both in terms of patterns of responding and in the confidence-accuracy relationship. Perhaps not surprisingly, the adults did not display any meaningful differences by lineup creator. Adults' lineup responding (and reported confidence) may be less sensitive to the relative similarity between fillers and the suspect than children's responding. This interpretation aligns with Fitzgerald et al. (2014), who found that increases in lineup member similarity substantially reduced children's correct identifications but had little impact on adults. Perhaps, then, children's eyewitness responding is more vulnerable to the perceptual clustering of lineup members, underscoring the importance of considering developmental factors in lineup construction.

Given the low probability of children being referenced to select lineup stimuli for an adult eyewitness, this finding offers interesting theoretical information about how lineups may be systemically biased against children. So why this inconsistency? These null findings may indicate that the benefits of child-created stimuli observed in Experiments 1 and 2 may not be ubiquitous but, instead, depend on factors such as lineup construction methods and sample variability. Experiment 3 included a smaller sample size, had a smaller filler selection pool (50 versus 100 images per suspect) and had a slightly older child sample. These differences, combined with the fact that children rated the fillers as more similar to the suspect (see above discussion), likely contributed to these null results. Poor encoding conditions are documented to exacerbate the difficulty of homogeneous lineups, reducing discriminability and increasing error rates (Clark 2003; Fitzgerald et al. 2014; Wells and Olson 2003). Thus, the combination of weak memory traces and tightly clustered similarity ratings likely contributed to the variability observed across experiments. It is also worth highlighting that the sensitivity analysis run for each experiment (see Supporting Information: Section 2.0) suggests that modest interaction effects could have gone undetected with the present sample; however, the opposite could also be true—with limited power, the probability of false discovery for interaction terms is higher. Therefore, Null or small interaction estimates should be interpreted as reflecting limited sensitivity, not definitive evidence of no effect. Of course, the lack of consistency in these results could also speak to our (generally) low understanding and inconsistent operationalization of similarity. The perceived similarity is just that—perceptions which are subjective and may be arbitrary in nature. Given how little we know about how similarity is judged, it is also possible that understanding of what "similar" means develops with time. We emphasize that our conclusions rely on relative patterns in similarity (i.e., which faces are judged closer/farther), not on strict equivalence of numeric scale use across ages. To minimize scale-use artifacts, we employed brief practice trials, age-appropriate anchors, and standardized instructions; nonetheless, factors such as numeracy and scale habits could have added noise that perhaps should be examined as a boundary condition for interpretation (cf. Valentine 1991; Jenkins et al. 2011). Facial processing and attention during similarity ratings were not the focus of this research, and, therefore, exploring how children judge similarity remains an interesting question for future research. Future explorations into which features children rely on while rating similarity could provide further insight into why child-created lineups might produce an advantage for child witnesses.

While our findings do not suggest that children should start selecting fillers in applied forensic settings (i.e., typically a police lineup operator assembles lineups; it would likely be problematic and/or unethical to ask children to be involved in real cases), they do challenge the assumption that a one-size-fits-all approach to lineup construction is neutral. If child witnesses are asked to make decisions based on stimuli constructed using adult perceptions of similarity, then researchers may be underestimating children's true memory abilities. This may partially account for the longstanding observation that children frequently underperform on lineup tasks compared to adults (Fitzgerald and Price 2015).

Interestingly, in Experiments 1 and 3 we found that children were more accurate in target-absent than target-present conditions—a reversal of the pattern typically observed in eyewitness research, particularly with children (Fitzgerald and Price 2015). This effect was not observed in Experiment 2, which used a live encoding interaction, suggesting that encoding quality may play a role. One possibility is that weaker memory traces from video exposure led children to adopt a more conservative strategy, producing higher rejection rates in target-absent conditions. Another possibility is that video encoding influenced attention or engagement differently than live encoding, reducing the likelihood of correct identifications in target-present lineups. Future work should examine how encoding modality (video vs. live) interacts with lineup construction and age to shape children's identification performance.

Another factor that may have contributed to the observed patterns is the potential influence of cross-group effects. Children may be particularly sensitive to own-age faces, showing better discrimination for same-age individuals than for adults (Rhodes and Anastasi 2012). Adults, conversely, may rely on broader configural strategies that are less sensitive to age, which could partly explain differences across samples. Likewise, all lineup stimuli in the present research were White, raising the possibility of cross-race effects for participants from other backgrounds, as recognition accuracy is reliably reduced when faces are from a different race than the witness (Meissner and Brigham 2001). These factors highlight the need for future work to examine how cross-age and cross-race influences interact with lineup construction methods to shape eyewitness identification accuracy.

This work adds to an existing body of work that argues that this age gap in lineup identification performance may, at least in part, be due to the context of a lineup task (Hiller and Weber 2013). Rather than attributing these age differences solely to cognitive immaturity, we should also scrutinize the methodological artifacts introduced by adult-centered lineup construction. If the goal is to understand the mechanisms that underlie this performance age gap in face recognition tasks, the lineup stimuli presented to children need to be a stronger focus. We cannot seek to understand these mechanisms using a context that may inherently place children at a disadvantage.

19 | Limitations and Future Research

This series of experiments represents a first attempt to explore an important methodological question: Does the age of the lineup stimulus creator influence children's lineup performance? Despite the evidence presented, these experiments are not without limitations. First, we tested this question using only six different face stimuli. To generalize these findings, future research should continue to explore this question using a broader sample of face stimuli. Second, it is important to highlight that we examined our research questions using lineups where the 'innocent suspect' in our target-absent conditions was the filler rated the most similar in appearance to the 'guilty suspect' (e.g., Navon 1992; Wells and Penrod 2011). This was done to reflect methods used in past developmental research (e.g., Zajac and Karageorge 2009). However, there are limitations associated with this method and future researchers interested in these questions should consider including moderately similar innocent suspects. For instance, using a median similarity innocent suspect may be a good approach to take with future research focusing on guilty–innocent suspect similarities (e.g., see Colloff et al. 2021). Third, the exploratory meta-analytic integration revealed moderate heterogeneity across experiments, suggesting that lineup construction effects may be sensitive to factors not systematically examined in this work, such as sample characteristics or encoding conditions. A great deal of lineup research uses video stimuli and, as such, future work should focus more on understanding how different encoding conditions influence memory.

An important limitation to consider is the possibility of participant fatigue or response patterns developing over the course of rating many faces. Although the task was kept brief (~10 min) and presentation order was randomized, future research should explore additional safeguards, such as intermittent breaks or alternative similarity tasks that may be more developmentally appropriate for younger children. In the present work, similarity ratings were collected using a 10-point Likert scale, which, although widely used in lineup construction research, may be less sensitive to subtle developmental differences in perception. In future work, alternative approaches such as rank-ordering or forced-choice tasks may provide additional insights into how children perceive facial similarity.

Another limitation of the present research is that our design was relatively complex, with multiple between-subjects factors, and our sample sizes were not planned to specifically optimize power for detecting higher-order interactions. Retrospective sensitivity analyses indicated that we had sufficient power to detect small-to-medium main effects, but considerably less sensitivity to detect interactions of comparable size. Thus, some null interactions may reflect limited power rather than the true absence of effects (Type II error). Moreover, when power is limited, the risk of Type I error for detecting these interactions also increases. Together, these considerations underscore the need to interpret any significant interactions as provisional, not definitive. Accordingly, these null findings should be interpreted with caution, and future studies with larger samples or simplified designs will be needed to more definitively assess the presence of interaction effects.

Additionally, it is worth noting that the adult sample in Experiment 3 was more ethnically diverse than the child samples, while the lineup stimuli were Caucasian. Research on the cross-race effect indicates that recognition accuracy can be reduced when witnesses view faces from a race different from their own (Meissner and Brigham 2001). This difference in sample composition may have influenced our findings, particularly if some adults were more likely to experience cross-race recognition challenges. By contrast, the child samples were less diverse and more ethnically aligned with the lineup stimuli, potentially reducing such effects. Future research should systematically examine how demographic characteristics, including race/ethnicity of both witnesses and lineup stimuli, interact with lineup construction methods to influence identification accuracy.

20 | Conclusion

The present work advances eyewitness science by testing whether developmentally appropriate stimuli (i.e., lineups created using children's perceived similarity ratings) influence the pattern of responding on lineups. The findings suggest that such stimuli can enhance children's performance, but further replication is needed, given the variability in how consistently this was observed. Importantly, these results challenge developmental researchers to critically consider methods used in lineup construction and warn us that not doing so may lead to a systematic underestimation of children's ability to perform on lineup tasks.

Author Contributions

Kaila C. Bruer: conceptualization, investigation, funding acquisition, writing – original draft, methodology, visualization, data curation, formal analysis, project administration, software, validation, writing

– review and editing. **Heather L. Price:** conceptualization, investigation, funding acquisition, writing – review and editing, methodology, data curation, supervision, project administration.

Acknowledgments

The authors thank Dr. Matthew Pechey, Jessica Williams, Dr. Madison Harvey, Alyssa Adams, Amanda Oliver, and Jordan Nixdorf for assistance with the data collection. We are sincerely appreciative of the Educating Youth in Engineering and Science summer camp at the University of Regina, and the parents and children for their support of this research.

Funding

This research was supported by a Social Science and Humanities Research Council (SSHRC) scholarship to the Kaila C. Bruer as well as a Natural Sciences and Engineering Research Council Discovery Development Grant to the Heather L. Price.

Ethics Statement

This research was approved by the REB at both the University of Regina and Thompson Rivers University and has closely followed the ethics guidelines put forth in the Tri-Council of Canada Policy Statement.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

[1] Reviewed peer-reviewed research papers are indicted with an asterisk (*) in the Reference section.

[2] Three research papers report using both describe and similarity methods and three papers did not clearly report their method for selecting fillers (Leippe et al. 1991; Lowenstein et al. 2010; Mertin 1989).

[3] $N = 9$ independent adult raters were asked to rate similarity for the female suspect and an additional 15 independent adult raters rated similarity for both female and male suspects in the originally conducted study.

References

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. "Mixed-Effects Modeling With Crossed Random Effects for Subjects and Items." *Journal of Memory and Language* 59, no. 4: 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67, no. 1: 1–48. https://doi.org/10.18637/jss.v067.i01.

*Beal, C. R., K. L. Schmitt, and D. J. Dekle. 1995. "Eyewitness Identification of Children: Effects of Absolute Judgments, Nonverbal Response Options, and Event Encoding." *Law and Human Behavior* 19: 197–216. https://doi.org/10.1007/BF01499325.

*Beresford, J., and M. Blades. 2006. "Children's Identification of Faces From Lineups: The Effects of Lineup Presentation and Instructions on Accuracy." *Journal of Applied Psychology* 91: 1102–1113. https://doi.org/10.1037/0021-9010.91.5.1102.

*Brewer, N., and K. Day. 2005. "The Confidence-Accuracy and Decision Latency-Accuracy Relationships in Children's Eyewitness Identification." *Psychiatry, Psychology and Law* 12: 119–128. https://doi.org/10.1375/pplt.2005.12.1.119.

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

Brown, A. L., and M. J. Kane. 1988. "Preschool Children Can Learn to Transfer: Learning to Learn and Learning From Example." *Cognitive Psychology* 20: 493–523. https://doi.org/10.1016/0010-0285(88)90014-X.

Bruce, V., and A. Young. 1998. *In the Eye of the Beholder: The Science of Face Perception*. Oxford University Press.

Bruer, K. C., R. J. Fitzgerald, H. L. Price, and J. D. Sauer. 2017. "How Sure Are You That This Is the Man You Saw? Using Confidence Judgments to Identify a Target With Child Eyewitnesses." *Law and Human Behavior* 41, no. 6: 541–555. https://doi.org/10.1037/lhb0000260.

Bruer, K. C., and H. L. Price. 2017. "A Repeated Forced-Choice Lineup Procedure Provides Suspect Bias Information With no Cost to Accuracy for Older Children and Adults." *Applied Cognitive Psychology* 31, no. 5: 448–466. https://doi.org/10.1002/acp.3342.

Burton, A. M., D. White, and A. McNeill. 2010. "The Glasgow Face Matching Test." *Behavior Research Methods* 42: 286–291. https://doi.org/10.3758/BRM.42.1.286.

Charman, S. D., and G. L. Wells. 2007. "Eyewitness Lineups: Is the Appearance-Change Instruction a Good Idea?" *Law and Human Behavior* 31, no. 1: 3–22. https://doi.org/10.1007/s10979-006-9006-3.

Clark, S. E. 2003. "A Memory and Decision Model for Eyewitness Identification." *Applied Cognitive Psychology* 17: 629–654. https://doi.org/10.1002/acp.891.

Clark, S. E. 2012. "Costs and Benefits of Eyewitness Identification Reform: Psychological Science and Public Policy." *Perspectives on Psychological Science* 7, no. 3: 238–259. https://doi.org/10.1177/1745691612439584.

Colloff, M. F., B. M. Wilson, T. M. Seale-Carlisle, and J. T. Wixted. 2021. "Optimizing the Selection of Fillers in Police Lineups." *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 8: e2017292118. https://doi.org/10.1073/pnas.2017292118.

*Davies, G., A. Tarrant, and R. Flin. 1989. "Close Encounters of the Witness Kind: Children's Memory for a Simulated Health Inspection." *British Journal of Psychology* 80: 415–429. https://doi.org/10.1111/j.2044-8295.1989.tb02333.x.

*Dehon, H., V. Vanootighem, and S. Brédart. 2013. "Verbal Overshadowing of Face Memory Does Occur in Children Too!" *Frontiers in Psychology* 4: 970. https://doi.org/10.3389/fpsyg.2013.00970.

*Dekle, D. J., C. R. Beal, R. Elliott, and D. Huneycutt. 1996. "Children as Witnesses: A Comparison of Lineup Versus Showup Identification Methods." *Applied Cognitive Psychology* 10: 1–12. https://doi.org/10.1002/(SICI)1099-0720(199602)10:1<1::AID-ACP354>3.0.CO;2-Y.

*Dunlevy, J. R., and J. Cherryman. 2013. "Target-Absent Eyewitness Identification Line-Ups: Why Do Children Like to Choose?" *Psychiatry, Psychology and Law* 20: 284–293. https://doi.org/10.1080/13218719.2012.671584.

Fitzgerald, R. J., and H. L. Price. 2015. "Eyewitness Identification Across the Lifespan: A Meta-Analysis of Age Differences." *Psychological Bulletin* 141: 1228–1265. https://doi.org/10.1037/bul0000013.

*Fitzgerald, R. J., H. L. Price, and D. A. Connolly. 2012. "Anxious and Nonanxious Children's Face Identification." *Applied Cognitive Psychology* 26: 585–593. https://doi.org/10.1002/acp.2833.

Fitzgerald, R. J., H. L. Price, C. Oriet, and S. D. Charman. 2013. "The Effect of Suspect-Filler Similarity on Eyewitness Identification Decisions: A Meta-Analysis." *Psychology, Public Policy, and Law* 19: 151–164. https://doi.org/10.1037/a0030618.

*Fitzgerald, R. J., B. F. Whiting, N. M. Therrien, and H. L. Price. 2014. “Lineup Member Similarity Effects on Children's Eyewitness Identification.” *Applied Cognitive Psychology* 28: 409–418. https://doi.org/10.1002/acp.3012.

Gao, Z., A. V. Flevaris, L. C. Robertson, and S. Bentin. 2011. “Priming Global and Local Processing of Composite Faces: Revisiting the Processing-Bias Effect on Face Perception.” *Attention, Perception, & Psychophysics* 73, no. 5: 1477–1486. https://doi.org/10.3758/s13414-011-0109-7.

Ge, L., G. Anzures, Z. Wang, et al. 2008. “An Inner Face Advantage in Children's Recognition of Familiar Peers.” *Journal of Experimental Child Psychology* 101: 124–136. https://doi.org/10.1016/j.jecp.2008.05.006.

*Goodman, G. S., J. E. Hirschman, D. Hepps, and L. Rudy. 1991. “Children's Memory for Stressful Events.” *Merrill-Palmer Quarterly (1982-)* 37, no. 1: 109–157. http://www.jstor.org/stable/23087341?seq=1#page_scan_tab_contents.

*Goodman, G. S., and R. S. Reed. 1986. “Age Differences in Eyewitness Testimony.” *Law and Human Behavior* 10: 317–332. https://doi.org/10.1007/BF01047344.

*Gross, J., and H. Hayne. 1996. “Eyewitness Identification by 5-To 6-Year-Old Children.” *Law and Human Behavior* 20: 359–373. https://doi.org/10.1007/BF01499028.

*Hafstad, G. S., A. Memon, and R. Logie. 2004. “Post-Identification Feedback, Confidence and Recollections of Witnessing Conditions in Child Witnesses.” *Applied Cognitive Psychology* 18: 901–912. https://doi.org/10.1002/acp.1037.

*Havard, C., and A. Memon. 2013. “The Mystery Man Can Help Reduce False Identification for Child Witnesses: Evidence From Video Line-Ups.” *Applied Cognitive Psychology* 27: 50–59. https://doi.org/10.1002/acp.2870.

*Havard, C., A. Memon, B. Clifford, and F. Gabbert. 2010. “A Comparison of Video and Static Photo Lineups With Child and Adolescent Witnesses.” *Applied Cognitive Psychology* 24: 1209–1221. https://doi.org/10.1002/acp.1645.

*Havard, C., A. Memon, P. Laybourn, and C. Cunningham. 2012. “Own-Age Bias in Video Lineups: A Comparison Between Children and Adults.” *Psychology, Crime & Law* 18: 929–944. https://doi.org/10.1080/1068316X.2011.598156.

Hiller, R. M., and N. Weber. 2013. “A Comparison of Adults' and Children's Metacognition for Yes/no Recognition Decisions.” *Journal of Applied Research in Memory and Cognition* 2, no. 3: 185–191. https://doi.org/10.1016/j.jarmac.2013.07.001.

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

Jenkins, R., D. White, X. Van Montfort, and A. M. Burton. 2011. “Variability in Photos of the Same Face.” *Cognition* 121, no. 3: 313–323. https://doi.org/10.1016/j.cognition.2011.08.001.

*Karageorge, A., and R. Zajac. 2011. “Exploring the Effects of Age and Delay on Children's Person Identifications: Verbal Descriptions, Lineup Performance, and the Influence of Wildcards.” *British Journal of Psychology* 102: 161–183. https://doi.org/10.1348/000712610X507902.

*Keast, A., N. Brewer, and G. L. Wells. 2007. “Children's Metacognitive Judgments in an Eyewitness Identification Task.” *Journal of Experimental Child Psychology* 97: 286–314. https://doi.org/10.1016/j.jecp.2007.01.007.

*Leippe, M. R., A. Romanczyk, and A. P. Manion. 1991. “Eyewitness Memory for a Touching Experience: Accuracy Differences Between Child and Adult Witnesses.” *Journal of Applied Psychology* 76: 367–379. https://doi.org/10.1037/0021-9010.76.3.367.

*Lindsay, R. C., J. D. Pozzulo, W. Craig, K. Lee, and S. Corber. 1997. “Simultaneous Lineups, Sequential Lineups, and Showups: Eyewitness Identification Decisions of Adults and Children.” *Law and Human Behavior* 21: 391–404. https://doi.org/10.1023/A:1024807202926.

Lindsay, R. C. L., and G. L. Wells. 1980. “What Price Justice? Exploring the Relationship of Lineup Fairness to Identification Accuracy.” *Law and Human Behavior* 4: 303–313. https://doi.org/10.1007/BF01040622.

*Lowenstein, J. A., H. Blank, and J. D. Sauer. 2010. “Uniforms Affect the Accuracy of Children's Eyewitness Identification Decisions.” *Journal of Investigative Psychology and Offender Profiling* 7: 59–73. https://doi.org/10.1002/jip.104.

Luus, C. A., and G. L. Wells. 1991. “Eyewitness Identification and the Selection of Distracters for Lineups.” *Law and Human Behavior* 15: 43–57. https://doi.org/10.1007/BF01044829.

Malpass, R. S., C. G. Tredoux, and D. McQuiston-Surrett. 2007. “Lineup Construction and Lineup Fairness.” In *The Handbook of Eyewitness Psychology, Vol II: Memory for People*, edited by R. C. L. Lindsay, D. F. Ross, J. D. Read, and M. P. Toglia, 155–178. Lawrence Erlbaum.

Mansour, J. K., J. L. Beaudry, N. Kalmet, M. I. Bertrand, and R. C. L. Lindsay. 2017. “Evaluating Lineup Fairness: Variations Across Methods and Measures.” *Law and Human Behavior* 41, no. 1: 103–115. https://doi.org/10.1037/lhb0000203.

Marin, B. V., D. L. Holmes, M. Guth, and P. Kovac. 1979. “The Potential of Children as Eyewitnesses.” *Law and Human Behavior* 3: 295–306. https://doi.org/10.1007/BF01039808.

Maurer, D., R. Le Grand, and C. J. Mondloch. 2002. “The Many Faces of Configural Processing.” *Trends in Cognitive Sciences* 6, no. 6: 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4.

Meissner, C. A., and J. C. Brigham. 2001. “Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review.” *Psychology, Public Policy, and Law* 7, no. 1: 3–35.

*Memon, A., and R. Rose. 2002. “Identification Abilities of Children: Does a Verbal Description Hurt Face Recognition?” *Psychology, Crime & Law* 8: 229–242. https://doi.org/10.1080/10683160208401817.

*Mertin, P. 1989. “The Memory of Young Children for Eyewitness Events.” *Australian Journal of Social Issues* 24: 23–32. https://doi.org/10.1002/j.1839-4655.1989.tb00854.x.

Mickes, L., M. B. Moreland, S. E. Clark, and J. T. Wixted. 2014. “ROC Analysis in Eyewitness Identification: Signal Detection and the Lure of the Lure.” *Journal of Applied Research in Memory and Cognition* 3, no. 2: 93–102.

Mondloch, C. J., S. Geldart, D. Maurer, and R. Le Grand. 2003. “Developmental Changes in Face Processing Skills.” *Journal of Experimental Child Psychology* 86, no. 1: 67–84. https://doi.org/10.1016/S0022-0965(03)00102-4.

Mondloch, C. J., R. Le Grand, and D. Maurer. 2002. “Configural Face Processing Develops More Slowly Than Featural Face Processing.” *Perception* 31: 553–566. https://doi.org/10.1068/p3339.

Mondloch, C. J., T. L. Lewis, D. R. Budreau, et al. 1999. “Face Perception During Early Infancy.” *Psychological Science* 10: 419–422. https://doi.org/10.1111/1467-9280.00179.

Navon, D. 1992. “Selection of Lineup Foils by Similarity to the Suspect Is Likely to Misfire.” *Law and Human Behavior* 16: 575–593. https://doi.org/10.1007/BF01044624.

*Parker, J. F., and L. E. Carranza. 1989. “Eyewitness Testimony of Children in Target-Present and Target-Absent Lineups.” *Law and Human Behavior* 13: 133–149. https://doi.org/10.1007/BF01055920.

*Parker, J. F., E. Haverfield, and S. Baker-Thomas. 1986. “Eyewitness Testimony of Children.” *Journal of Applied Social Psychology* 16: 287–302. https://doi.org/10.1111/j.1559-1816.1986.tb01141.x.

*Parker, J. F., and V. Ryan. 1993. “An Attempt to Reduce Guessing Behavior in Children's and Adults' Eyewitness Identifications.” *Law and Human Behavior* 17: 11–26. https://doi.org/10.1007/BF01044534.

Police Executive Research Forum. 2013. “A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies (NCJ 242617).” https://nij.ojp.gov/library/publications/national-survey-eyewitness-identification-procedures-law-enforcement-agencies.

*Pozzulo, J. D., and J. Balfour. 2006. “Children's and Adults' Eyewitness Identification Accuracy When a Culprit Changes His Appearance: Comparing Simultaneous and Elimination Lineup Procedures.” *Legal and Criminological Psychology* 11: 25–34. https://doi.org/10.1348/135532505X52626.

*Pozzulo, J. D., and J. Dempsey. 2006. “Biased Lineup Instructions: Examining the Effect of Pressure on Children's and Adults' Eyewitness Identification Accuracy.” *Journal of Applied Social Psychology* 36: 1381–1394. https://doi.org/10.1111/j.0021-9029.2006.00064.x.

*Pozzulo, J. D., J. Dempsey, K. Bruer, and C. Sheahan. 2012. “The Culprit in Target-Absent Lineups: Understanding Young Children's False Positive Responding.” *Journal of Police and Criminal Psychology* 27: 55–62. https://doi.org/10.1007/s11896-011-9089-8.

*Pozzulo, J. D., J. Dempsey, S. Corey, A. Girardi, A. Lawandi, and C. Aston. 2008. “Can a Lineup Procedure Designed for Child Witnesses Work for Adults? Comparing Simultaneous, Sequential, and Elimination Lineup Procedures.” *Journal of Applied Social Psychology* 38: 2195–2209. https://doi.org/10.1111/j.1559-1816.2008.00387.x.

*Pozzulo, J. D., J. Dempsey, and C. Crescini. 2009. “Preschoolers' Person Description and Identification Accuracy: A Comparison of the Simultaneous and Elimination Lineup Procedures.” *Journal of Applied Developmental Psychology* 30: 667–676. https://doi.org/10.1016/j.appdev.2009.01.004.

*Pozzulo, J. D., J. L. Dempsey, C. Crescini, and J. M. Lemieux. 2009. “Examining the Relation Between Eyewitness Recall and Recognition for Children and Adults.” *Psychology, Crime & Law* 15: 409–424. https://doi.org/10.1080/10683160802279625.

*Pozzulo, J. D., J. L. Dempsey, and K. Wells. 2010. “Does Lineup Size Matter With Child Witnesses.” *Journal of Police and Criminal Psychology* 25: 22–26. https://doi.org/10.1007/s11896-009-9055-x.

*Pozzulo, J. D., and R. C. L. Lindsay. 1999. “Elimination Lineups: An Improved Identification Procedure for Child Eyewitnesses.” *Journal of Applied Psychology* 84: 167–176. https://doi.org/10.1037/0021-9010.84.2.167.

*Price, H. L., and R. J. Fitzgerald. 2016. “Face-Off: A New Identification Procedure for Child Eyewitnesses.” *Journal of Experimental Psychology: Applied* 22, no. 3: 366–380.

Price, H. L., K. C. Bruer, and M. Adkins. 2020. “Using an Interactive Simultaneous Lineup Procedure Can Identify Looking Behaviour That Predicts Accuracy in Children and Adult Eyewitnesses.” *Law and Human Behavior* 44, no. 3: 223–237. https://doi.org/10.1037/lhb0000364.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rhodes, M. G., and J. S. Anastasi. 2012. “The Own-Age Bias in Face Recognition: A Meta-Analytic and Theoretical Review.” *Psychological Bulletin* 138, no. 1: 146–176. https://doi.org/10.1037/a0025750.

*Ross, D. F., D. F. Marsil, T. R. Benton, et al. 2006. “Children's Susceptibility to Misidentifying a Familiar Bystander From a Lineup: When Younger Is Better.” *Law and Human Behavior* 30: 249–257. https://doi.org/10.1007/s10979-006-9034-z.

Searcy, J. H., and J. C. Bartlett. 1996. “Inversion and Processing of Component and Spatial–Relational Information in Faces.” *Journal of Experimental Psychology: Human Perception and Performance* 22, no. 4: 904–915.

Taylor, M. J., G. E. Edmonds, G. McCarthy, and T. Allison. 2001. “Eyes First! Eye Processing Develops Before Face Processing in Children.” *Neuroreport* 12: 1671–1676. https://www.ncbi.nlm.nih.gov/pubmed/11409737.

Valentine, T. 1991. “A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition.” *Quarterly Journal of Experimental Psychology Section A* 43, no. 2: 161–204. https://doi.org/10.1080/14640749108400966.

Want, S. C., O. Pascalis, M. Coleman, and M. Blades. 2003. “Recognizing People From the Inner or Outer Parts of Their Faces: Developmental Data Concerning “Unfamiliar” Faces.” *British Journal of Developmental Psychology* 21: 125–135. https://doi.org/10.1348/026151003321164663.

Wells, G. L., M. B. Kovera, A. B. Douglass, N. Brewer, C. A. Meissner, and J. T. Wixted. 2020. “Policy and Procedure Recommendations for the Collection and Preservation of Eyewitness Identification Evidence.” *Law and Human Behavior* 44, no. 1: 3–36. https://doi.org/10.1037/lhb0000359.

Wells, G. L., and E. A. Olson. 2003. “Eyewitness Testimony.” *Annual Review of Psychology* 54, no. 1: 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028.

Wells, G. L., and S. D. Penrod. 2011. “Eyewitness Identification Research: Strengths and Weaknesses of Alternative Methods.” In *Research Methods in Forensic Psychology*, edited by B. Rosenfeld and S. D. Penrod, 237–256. Wiley.

Wells, G. L., S. M. Rydell, and E. P. Seelau. 1993. “The Selection of Distractors for Eyewitness Lineups.” *Journal of Applied Psychology* 78: 835–844. https://doi.org/10.1037/0021-9010.78.5.835.

Wilson, R. R., M. Blades, and O. Pascalis. 2007. “What Do Children Look at in an Adult Face With Which They Are Personally Familiar?” *British Journal of Developmental Psychology* 25: 375–382. https://doi.org/10.1348/026151006X159112.

Wogalter, M. S., R. S. Malpass, and D. E. Mcquiston. 2004. “A National Survey of US Police on Preparation and Conduct of Identification Lineups.” *Psychology, Crime & Law* 10: 69–82. https://doi.org/10.1080/10683160410001641873.

Wogalter, M. S., D. B. Marwitz, and D. C. Leonard. 1992. “Suggestiveness in Photospread Line-Ups: Similarity Induces Distinctiveness.” *Applied Cognitive Psychology* 6: 443–453. https://doi.org/10.1002/acp.2350060508.

*Zajac, R., and A. Karageorge. 2009. “The Wildcard: A Simple Technique for Improving Children's Target-Absent Lineup Performance.” *Applied Cognitive Psychology* 23: 358. https://doi.org/10.1002/acp.1511.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** acp70149-sup-0001-TableS1-S10.docx.