How sure are you that this is the man you saw? Child witnesses can use confidence judgments to identify a target

Kaila C. Bruer[1], Ryan J. Fitzgerald[2], Heather L. Price[3], and James D. Sauer[4]

[1]University of Regina

[2]University of Portsmouth

[3]Thompson Rivers University

[4]University of Tasmania

This article is published in *Law and Human Behavior*.

This research was supported by a Natural Sciences and Engineering Research Council Discovery Development Grant to the Heather L. Price. 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.

Correspondence concerning this article should be addressed to Kaila C. Bruer, Department of Psychology, University of Regina, Administration-Humanities Building, AH 345, 3737 Wascana Parkway, Regina, SK, S4S 0A2. E-mail: bruer20k@uregina.ca

Abstract

We tested whether an alternative lineup procedure designed to minimize problematic influences (e.g., metacognitive development) on decision criteria could be effectively used by children and improve child eyewitness identification performance relative to a standard identification task. 516 children (6- to 13-year-olds) watched a video of a target reading word lists and, the next day, made confidence ratings for each lineup member or standard categorical decisions for 8 lineup members presented sequentially. Two algorithms were applied to classify confidence ratings into categorical decisions and facilitate comparisons across conditions. The classification algorithms produced accuracy rates for the confidence rating procedure that were comparable to the categorical procedure. These findings demonstrate that children can use a ratings-based procedure to discriminate between previously seen and unseen faces. In turn, this invites more nuanced and empirical consideration of ratings-based identification evidence as a probabilistic index of guilt that may attenuate problematic social influences on child witnesses' decision criteria.

Key words: child, eyewitness, confidence judgments, lineup identification

Public Significance Statement

Child eyewitnesses are prone to choosing incorrectly from traditional identification lineups. This research demonstrates that children's confidence ratings can provide meaningful information about the quality of their memories for faces and the degree to which they recognize previously seen or unseen faces presented in identification lineups.

How sure are you that this is the man you saw? Child witnesses can use confidence judgments to identify a target

Even in the most ideal situation eyewitness identifications can be inaccurate (Wells & Olson, 2003)—this is especially true for child eyewitnesses who are more likely than adult eyewitnesses to identify an innocent person from a perpetrator-absent lineup (Fitzgerald & Price, 2015). Given the fallibility of eyewitness memory, the approaches traditionally used to administer lineups to witnesses have been scrutinized (Brewer & Wells, 2011; Wells, Memon, & Penrod, 2006). In response to this scrutiny, an alternative approach to improving accuracy with adult eyewitnesses was developed to mitigate factors that may influence witnesses' decision criteria and increase error rates (Brewer, Weber, Wootton, & Lindsay, 2012; Sauer, Brewer, & Weber, 2008; Weber & Varga, 2012). The alternative approach permits eyewitnesses to provide a confidence judgment for each lineup member (reflecting their likelihood of guilt), rather than a traditional categorical decision. An algorithm that uses the distribution of confidence ratings can then be applied to derive identification and rejection classifications. This procedure has been effective at increasing accuracy for adult witnesses, particularly for perpetrator-absent lineups.

Child eyewitnesses, however, present a unique problem to the legal system. Research consistently demonstrates that child eyewitnesses are prone to choosing incorrectly from a lineup –especially the youngest children studied, those aged 5-8 years (Fitzgerald & Price, 2015). Because of their tendency to choose, children are particularly challenged when the perpetrator is absent from the lineup (Fitzgerald & Price, 2015; Pozzulo & Lindsay, 1998). Children's problematic choosing may reflect the setting of overly-lenient decision criteria (i.e., low threshold for selecting a lineup member) that results from peripheral factors, such as implicit social pressure to choose (Pozzulo, Dempsey, Bruer, & Sheahan, 2012). However, research has

yet to examine whether confidence ratings – a procedure that avoids single, explicit categorical decisions, potentially reducing the impact of non-diagnostic influences on criterion placement – can be used by children to effectively identify a target among foils in a lineup. We explored whether using confidence ratings could improve child eyewitness identification performance, relative to a standard identification task.

Confidence Ratings as Indices of Memory

The recognition memory literature has established a strong link between confidence and accuracy. Decision theories of recognition, including signal detection theory, generally posit that confidence represents the degree of match between a stimulus and an image in memory (Green & Swets, 1966; Leippe, 1980; Wickelgren & Norman, 1966). Similarly, evidence accumulator models propose that confidence represents the difference between the evidence that an item has been seen and the evidence that an item is new (Vickers, 1979). Viewing a previously seen item should create a stronger connection to memory than viewing a never-before-seen item. As a result, confidence tends to increase with accuracy (Norman & Wickelgren, 1965; Trow, 1923; Van Zandt, 2000).

There is a long history of obtaining a confidence judgment as part of the eyewitness identification paradigm. Confidence judgments obtained immediately after the identification decision can be informative about likely accuracy (Brewer & Weber, 2008; Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996), provided that the information has been processed under favorable conditions (Deffenbacher, 1980), a positive identification has been made (Sporer, Penrod, Read, & Cutler, 1995), and no administrator feedback has been given (Wells & Bradfield, 1998). To be clear, an expression of high confidence in an identification decision is by no means conclusive evidence that the decision was accurate. Confident witnesses can be wrong.

However, when aggregated across individuals, a relation between confidence and accuracy is typically found, particularly if calibration analyses are performed (Brewer & Wells, 2006; Juslin et al., 1996).

As mentioned above, a new use of confidence has recently emerged in the eyewitness identification literature. The conventional role of confidence ratings in eyewitness procedures has been to supplement a categorical lineup decision with a single post-identification confidence rating (e.g., how confident are you in your final decision?). Rather than asking witnesses to rate their confidence in an identification decision, Sauer and colleagues (2008) asked participants to provide a confidence rating for each lineup member, indexing the likelihood that the lineup member in question was the perpetrator (without making a categorical identification). Their objective was to minimize problematic influences on decision criteria (e.g., demand characteristics). An additional benefit of this method was that witnesses made confidence decisions for each lineup member, rather than using all the information in the full lineup to make a single categorical decision followed by a single confidence rating. Thus, the confidence procedure reduced the amount of information that witnesses had to sort through into more manageable judgments and provided a richer source of information about the extent to which individual lineup members matched witnesses' memory for the culprit.

Sauer et al. (2008) applied an algorithm to derive a positive (i.e., choosing or making an identification) or negative (i.e., not choosing or rejecting) classification from the witnesses' confidence judgments. Following the application of the algorithm, the confidence procedure yielded accuracy rates comparable to traditional, categorical decisions when the target was present and provided a considerable advantage when the target was absent. These results suggest that confidence ratings can be used by adult witnesses to accurately discriminate previously seen

from unseen faces (i.e., confidence served as an index of recognition memory). Subsequent research has confirmed that adults possess the metacognitive ability to give confidence ratings that indicate the degree of match between each picture and their memory of the perpetrator (Brewer et al., 2012; Sauer et al., 2012). Whether children are able to use confidence ratings similarly remains an empirical question.

Children as Eyewitnesses

The cause of children's propensity to choose is not fully understood, but it is likely the result of a convergence of memory, cognitive development, and social-influence factors (e.g., Beal, Schmitt, & Dekle, 1995; Pozzulo et al., 2012). There have been several attempts made to reduce children's choosing during lineups (e.g., Pozzulo & Lindsay, 1999; Price & Fitzgerald, 2016; Zajac & Karageorge, 2008). While these attempts have found some success, children's high rate of choosing continues to be a problem (see Fitzgerald & Price, 2015).

Age differences in choosing may be explained by children's use of overly-lenient decision criteria (Humphries, Holliday, & Flowe, 2012). The ability to monitor and regulate decision criteria is dependent on metacognitive abilities (Flavell & Wellman, 1977; Haller, Child, & Walberg, 1988) that develop through childhood and into adolescence (e.g., Bryce & Whitebread, 2012; Keast, Brewer, & Wells, 2007; Roebers, 2002). With limited ability to monitor and regulate their cognitive processes, children—especially children younger than 9 years old (e.g., Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001; Roebers & Howie, 2003)—may not be able to recognize memory limitations and adjust their decision criterion accordingly. Dunlevy and Cherryman (2013) argued that when a target is absent from a lineup, children (aged 6 to 7) adjust their decision criteria—but not in the desired direction. Instead of using more conservative criteria (high threshold when selecting a lineup member) when there is no close

match or when their memory for a perpetrator is weak, children appear to use more lenient decision criteria.

Children's use of lenient decision criteria is likely augmented by perceived pressure to choose and a desire to acquiesce. An eyewitness identification task involves a level of implicit pressure to pick someone. Participants may believe that choosing none of the lineup members gives the impression that they are unwilling to complete the task (Wells & Luus, 1990). Children (4- to 11-years-old) appear to be especially vulnerable to this type of social pressure due to exaggerated power/authority differences between a child and an interviewer (Beal et al., 1995; Parker & Ryan, 1993; Pozzulo et al., 2012). Thus, children may choose from lineups because they want to please their interviewer.

Given their propensity to choose, children might benefit from a procedure that encourages more conservative responding. Sequential presentation of lineup images has been demonstrated to make adults more conservative (e.g., Palmer & Brewer, 2012; Steblay, Dysart & Wells, 2011). However, presenting lineup members individually still results in high choosing rates in children (Lindsay, Pozzulo, Craig, Lee, & Corber, 1997; Parker & Ryan, 1993; Pozzulo & Lindsay, 1998). Children have also been found to choose more than one lineup member—likely due to the use of lenient decision criteria or trouble understanding the task. However, adjusting a lineup task to involve a number of smaller confidence decisions, rather than a single categorical decision, may help to minimize problematic influences on children's decision criteria.

Children and Confidence Ratings

Although a positive relation between confidence and accuracy has been demonstrated for adult witnesses (e.g., Juslin et al., 1996; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer,

Brewer, Zweck, & Weber, 2010), there is little evidence of a similar relation in children (Brewer & Day, 2005; Keast et al., 2007). Specifically, when children (10- to- 13-years old) pick from a lineup, they show greater overconfidence and poorer calibration (cf. adults; Keast et al., 2007). However, in previous lineup research with children, the task involved a retrospective judgment of confidence about a categorical identification. Findings in the developmental metacognitive literature suggest children may nevertheless be able to use confidence as an index of memory, thus suggesting the lineup literature has just not yet found how to make such a procedure work for child witnesses.

Basic metacognitive processing develops during pre-school years (Schneider & Lockl, 2008) and there is evidence that children as young as 8 years old can monitor the accuracy of retrieved memories and strategically regulate the reporting of memories to improve accuracy (e.g., Koriat et al., 2001; Roebers & Howie, 2003). Hiller and Weber (2013) recently explored the role of confidence and metacognitive development in a word-pair recognition task that required discrimination of seen from unseen stimuli (akin to a lineup task). Although children's (8- to 12-year-olds) confidence ratings were less well calibrated than adults' (i.e., the degree of correspondence between the level of confidence expressed for, and the probable accuracy of, a given response was weaker), higher confidence ratings were still associated with more accurate decisions. Importantly, despite children's overconfidence, similar levels of confidence-based discrimination were found between children and adults.

Instructing children to rate their confidence that each lineup member is the target may be an effective method to capitalize on the utility of confidence information. If children's higher false identification rates (cf. adults) stem from the nature of the traditional eyewitness task (Hiller & Weber, 2013), then using a method that circumvents a typical lineup decision may be

more reliable. With a confidence rating procedure, children are not required to make a single, categorical lineup identification and instead are only asked to rate their confidence as to whether each lineup member is the target. Contrary to a categorical task, a confidence rating procedure minimizes the need for children to consider the possibility that the target is not in the lineup. Children need only consider the relative likelihood that each lineup member is the target, which seems to be within the range of children's metacognitive abilities (Roebers & Howie, 2003).

A confidence rating procedure also changes the lineup task from a single decision involving numerous stimuli to a series of responses, each to a single stimulus, which may be particularly advantageous for children. Making a categorical lineup identification requires complex processing (i.e., assessing which one face matches their memory of the target better than other faces) that induces a large cognitive load and, in turn, may negatively impact performance (Hiller & Weber, 2013; Pozzulo & Lindsay, 1999). Circumventing the need for a child to make a categorical identification could reduce the cognitive load associated with the task, alleviate inherent pressure to choose that is associated with making a single, categorical identification (Beal et al., 1995; Price & Fitzgerald, 2016) and mitigate problems associated with use of overly-lenient decision criteria. Thus, children may be able to use confidence ratings to discriminate previously seen from unseen faces.

Present Study

The goals of this research were two-fold. Our primary goal was to assess the utility of a confidence paradigm with school-aged children. That is, can children use confidence ratings to accurately discriminate a previously seen face from previously unseen faces? Can children's confidence ratings be used as an index of recognition? Our second objective was to see how the accuracy of children's responses from the confidence procedure would compare to a categorical

procedure – in this case, a sequential categorical procedure. A sequential presentation style was used to compare categorical and confidence procedures because confidence ratings are most useful when they index the similarity of a lineup member to memory; sequential presentation reduces the possibility for relative similarity to interfere (cf. simultaneous presentation).

Children (aged 6-8 and 9-13 years) viewed a video of a target and then completed a categorical or confidence lineup procedure on the following day. For the confidence procedure, confidence ratings were collected for each lineup member and then classified as positive (those who made an identification) or negative (those who rejected the lineup) decisions (Koriat & Goldsmith, 1996; Weber & Varga, 2012). These classifications were then compared to responses from children who made categorical lineup decisions. Both the confidence and categorical procedure presented the lineup members sequentially. We hypothesized that: (1) children's confidence ratings could be used to discriminate between guilty and innocent suspects and (2) the confidence procedure would improve children's response accuracy compared to the categorical procedure. This improved accuracy was expected to be driven by a reduction in inaccurate positive classifications. Moreover, considering the substantive improvements in meta-cognitive abilities associated with development through this age group, we hypothesized that these anticipated improvements would be more pronounced for older than younger children.

Method

Participants and Design

We recruited 516 children, 6- to 13-years-old (M_{age} = 8.89, *SD* = 1.88; 57% males), from local camps and after-school programs. This study was approved by the University of Regina Research Ethics Board. During program drop-off times, researchers discussed the study with parents/guardians, who signed consent forms if interested. Children with parental consent and

who themselves assented were randomly assigned to one of four conditions in a 2 (response procedure: confidence, categorical) x 2 (target: present vs. absent) design. Due to the broad age range included in this study, we also examined age differences between younger (6- to 8-year-olds, M_{age} = 7.18, *SD* = 0.77; 54% males) and older children (9- to 13-year-olds, M_{age} = 10.29, *SD* = 1.29; 60% males). These age groupings were chosen for a few reasons. First, a recent meta-analysis reported that child eyewitnesses start performing more like adults on lineup tasks at around age 9 (Fitzgerald & Price, 2015). For the younger group, 6 through 8 years is a time period in which there is rapid change in cognitive development, particularly in meta-cognition. Meta-cognitive developments around the age of 9-10 years allow children to begin to perform at adult-like levels on some tasks (e.g., association/recall tasks; Schneider, 1986; Wellman, 1978). Thus, as children grow past 9 years of age, we can expect to observe a more gradual increase with an eventual plateauing of identification accuracy.

Materials

Target event and lineup stimuli. The video event and lineup stimuli were used in a previous study with children (Price & Fitzgerald, 2016). The video alternated between a woman performing magic tricks and a man reading word lists. The man reading word lists served as the target person. The video was just over 6 minutes long and, of that time, the target was in view for approximately 2 minutes. All lineups contained eight members, presented sequentially. Target-present lineups comprised the target and seven fillers. Target-absent lineups comprised a designated "innocent suspect" and the same fillers as in the target-present lineup.

Similarity ratings were obtained to guide selection of the innocent suspect and fillers. Independent raters (*n* = 35 adults) completed 200 trials in which the target person was presented alongside another person of the same race and sex. For each pair, raters were asked: "In terms of

physical appearance, how similar are these two individuals?" Ratings were completed on a 10-point scale, ranging from 1 (highly dissimilar) to 10 (highly similar). Ratings ranged from 1.49 to 6.06 (M = 3.51, SD = 0.78).

In applied settings, lineups are typically constructed by matching fillers to the appearance of the suspect (Police Executive Research Forum, 2013). For target-absent lineups, this procedure has been theorized to lead to the selection of fillers who would not resemble the culprit to same degree as would an innocent suspect (Navon, 1992). Accordingly, the designated innocent suspect was the person with the highest similarity rating (M = 6.06). The average similarity ratings for the fillers (M = 4.34, SD = 0.37) were some of the highest in the set, but were all lower than the average rating for the innocent suspect.

Confidence cup scale. Similar to previous research (e.g., Brewer & Day, 2005), children made confidence judgments using a 5-point scale containing numbers 0 through 4. In addition, each number was accompanied by a picture of a drinking glass with increasing levels of water in it (see Appendix A). A 5-point scale was used to avoid confusion stemming from a larger visual scale. For example, adjacent glasses in a 10-point scale would have had less noticeable differences in water. A 5-point scale allowed for a clear top, bottom, mid-point, and two half-way to mid-points (see Appendix A). Previous researchers reported success using a similar visual scale with children (the Cup Scale; Weston, Boxer, & Heatherington, 1998).

Procedure

On the first day, children watched the video containing the male target as a group. The following day, a team of research assistants interviewed the children individually to administer the lineup task. Research assistants were blind to the identity of the target. After a short rapport-building session, research assistants reminded the children about the target video and noted that

the target person may or may not be in the stack of pictures. Children were told they could take as long as they needed to look at each picture. Research assistants showed participants eight lineup members sequentially, in a random order. Children made a decision for each lineup member before proceeding to the next. All lineup identification procedures and decisions were audio-recorded to ensure the research assistants followed the protocol.

Confidence condition. For each picture, children used the confidence cup scale to indicate how sure they were that the picture depicted the target. Verbal instructions were accompanied by gestures to appropriate cups. Below are the instructions provided to the children:

> *I don't know what Jordan looks like, so I need you to help me figure out if his picture is in this pile. There might be a picture of Jordan in this pile or there might not be a picture of Jordan in this pile. Now I'm going to show you pictures one at a time. You can look at the picture for as long as you want. For each picture, I am going to ask you how sure you are that it is Jordan. Here are some pictures of cups with different amounts of water in them. I want you to tell me how sure you are by picking a cup with the right amount of water in it. It works like this: The more sure you are that the person is Jordan, the more water will be in the cup. If you are not really sure that it is Jordan, choose a cup that doesn't have very much or any water in it. If you are a little bit sure but not too sure that it is Jordan, you should choose a cup that has some water in it but it shouldn't be totally full. If you are really sure that it is Jordan, choose a cup that is almost or totally full. Does that all make sense? How sure are you that this is Jordan?*

Confidence Scale Check. In an initial check of children's ability to use the confidence scale, a pilot study was conducted with ten children (M_{age} = 11.00) and children's responses

indicated ability to understand and apply the scale. For an additional 334 participants who completed the full study, following the lineup task, we presented children with a similar visual analog scale (jars filled with jelly beans) and asked children to show which end of the scale indicated being not at all sure (and very sure). Children were then asked to place the jars in order of increasing confidence. For all children, this procedure took place following the lineup task and a brief distracter task and, thus, had no impact on their performance with the Cup Scale. Of those who completed the jelly bean task, 93% of children completed the task with no errors or guidance, while 7% had minor difficulty but could complete it with guidance.

In addition, there was some concern that children would use the cup scale in a dichotomous or binary way (i.e., primarily selecting from either end of the scale). However, children used all response options on the cup scale and the distribution of responses at each possible option indicates willingness to spread across the scale, with an intuitively decreasing frequency with higher confidence ratings (younger children: 0 = 43%, 1 = 21%, 2 = 18%, 3 = 10%, 4 = 8%; older children: 0 = 44%, 1 = 24%, 2 = 19%, 3 = 9%, 4 = 4%).

Categorical condition. Participants in the categorical condition were instructed to provide a traditional yes/no decision indicating whether each picture depicted the man from the video:

> *I don't know what Jordan looks like, so I need you to help me figure out if his picture is in this pile. There might be a picture of Jordan in this pile or there might not be a picture of Jordan in this pile. I'm going to show you pictures one at a time. You can look at the picture for as long as you want. For each picture, I'm going to ask, "Is this Jordan?" If it's Jordan's picture, say "yes". Remember though, Jordan's picture might not be in the pile. If it's not Jordan's picture, say "no". There's one last thing you should know before*

we get started. I want to ask you about all of the pictures, so even if you've already told me that one of the pictures is Jordan I'm going to keep asking you "Is this Jordan" until you've seen all the pictures. Does that all make sense? Is this Jordan?

After each decision, children provided a post-identification confidence assessment using the same cup scale as in the confidence procedure condition. We compared the classification accuracy of those children who only made a confidence rating in the confidence procedure with children who made a post-identification confidence rating in the categorical procedure. For both age groups, we found evidence to support asking for confidence ratings alone rather than following a categorical decision (see Online Supplementary Materials for further discussion). Moreover, when a confidence rating was provided following a categorical decision, the response classification of the confidence ratings only mapped onto the categorical decisions in less than 50% of the cases (45% for younger children, 49% for older children).

After children viewed all pictures in the lineup, the next step depended on the number of 'yes' decisions made, if any. If a single picture was selected, the procedure ended. If multiple selections were made, (n = 27 or 24% for younger children, n = 30 or 21% for older children) this was resolved by repeating the procedure and stopping after the first positive identification. Lastly, if children answered "no" to all lineup members, they were asked to provide an overall assessment of confidence to indicate how sure they were that the target was not shown; however, these results were not a focus of the present analyses.

Classification Methods

To determine classification accuracy for the confidence condition, confidence ratings were evaluated against a criterion, or critical value, that produced an estimate of the proportion of target-present trials that best matched the actual proportion of target-present trials (Koriat &

Goldsmith, 1996; Weber & Varga, 2012). This contrasts with Sauer and colleagues' (2008; 2012) approach, which maximized the proportion of correct decisions. However, as discussed by Weber and Varga (2012), the algorithms used by Sauer and colleagues may result in an artificial inflation of classification accuracy. To provide an unbiased comparison between lineup procedures, Weber and Varga optimized the data to closely match the designed proportion of target presence.

We used Weber and Varga's (2012) method by applying four previously-used classification algorithms to the children's confidence ratings (Sauer et al., 2008, 2012). Specifically, we examined the C1 (i.e., MAX ONLY), C2 (i.e., MAX vs. NEXT), C3 (i.e., MAX vs. AVERAGE), and hierarchical classification methods. However, all examined classification methods produced similar results. For parsimony and to allow for comparison with previous research (e.g., Sauer et al., 2008), we report only the results of the C3 and hierarchical classification methods. For a complete description of the four classification methods see the Online Supplementary Materials.

The Solver add-in function in *Microsoft Excel* (2010) was used to optimize (using the evolutionary method) a classification method's criterion until the proportion of positive decisions classified deviated least from the proportion of target-present trials (see Weber & Varga, 2012). For example, given that 71 older children were assigned to a target-present lineup, we ran Solver until approximately the same number of responses was classified as positive. Each algorithm was maximized separately for each age group. Table 1 displays criterion scores for each algorithm.

Table 1

The identified criterion and optimized proportion of correct classifications for each method

		MAX vs. AVERAGE			
		(C3)	H1	H2	H
Younger	Criterion	1.73	0.20	2.05	
	Proportion of correct classifications	50.86%	62.90%	74.00%	67.86%
Older	Criterion	1.68	0.90	0.56	-
	Proportion of correct classifications	59.29%	75.64%	64.52%	70.71%

Note. H = hierarchical classification. The H proportion of correct classifications represents the average of the H1 and H2 proportions.

For each classification method, the associated criterion was used for categorizing confidence judgments as positive (identification) or negative (rejection) decisions. When using MAX vs. AVERAGE, the criterion represented the required difference between the picture that received the highest confidence rating and the average of all remaining confidence ratings. The hierarchical method (H) involved two steps. In the first step, a single criterion was calculated using only the confidence ratings for the suspect's picture (H1). Next, another criterion was calculated using the confidence judgments from the remaining seven fillers (H2). Suspect confidence ratings that reached the H1 criterion were classified as positive identifications. Filler confidence ratings that reached the H2 criterion were classified as negative identifications (i.e., rejection) because they were known errors. Any decision that did not reach the H1 or H2 criterion was considered to represent evidence too weak to be classified as either an identification or a rejection. Accordingly, these decisions were classified as 'indeterminate', which is conceptually similar to a "don't know" response. For all classification methods, we applied the conservative approach used by Sauer et al. (2008), such that participants who did not

assign a unique maximum confidence rating to one of the eight pictures (e.g., multiple maxes, all zero confidence) were considered to have rejected the lineup.

Results

To address our research questions, several analyses were performed. First, we explored different indices to examine children's ability to use confidence ratings. Next, we compared the classifications of children's responding in the confidence procedure with children's decisions in the categorical procedure. Specifically, we examined for differences in suspect identification responses and response accuracy across these procedures. To further compare performance across procedures, diagnosticity and discrimination analyses were performed. Lastly, we conducted a profile analysis to examine the utility of children's confidence ratings at an individual level.

Can confidence ratings provided by children be used as an index of recognition?

We first investigated whether participants could use confidence ratings to discriminate previously seen (target) from unseen faces (fillers and innocent suspect). This is among the most important questions in the current work because evidence of children's ability to effectively use confidence ratings to discriminate a target from unfamiliar faces would provide further justification for exploration of such a technique. The adjusted normalized discrimination index (ANDI) was calculated to provide a measure of how well participants' confidence ratings of each lineup member discriminated guilty from innocent suspects (for the formulae, see Yaniv, Yates, & Smith, 1991). ANDI is a measure of variance in accuracy accounted for by confidence ratings and it ranges from 0 (no discrimination) to 1 (perfect discrimination). A bootstrapping procedure was used to compute .05% inferential confidence intervals. This procedure (see Palmer, Brewer, & Weber, 2010; Tryon, 2001) used the observed data as a sampling distribution and conducted

3000 replications to estimate variance of ANDI. This estimated variance provided the distribution needed to calculate confidence intervals. The ANDI scores revealed that both younger (.20, .05 ICI [.19, .22]) and older (.24, .05 ICI [.23, .25]) children were able to use confidence to discriminate a target from unseen faces: 20% and 24% of the variance in outcomes was explained by confidence ratings for younger and older children, respectively.

As further evidence of how children used the confidence scale, we examined whether multiple confidence ratings provided by children could be used to accurately classify suspects as guilty or innocent. We calculated the proportion of correct classifications (correct identifications and correct rejections) using the classification methods described above (see Table 1). The high proportion of correct classifications suggest that children's confidence ratings can be used to effectively classify previously seen (i.e., target) and unseen faces (innocent suspect and fillers).

Accuracy in Target Present and Target Absent Lineups

Classification outcomes of the algorithms are presented in Table 2. To avoid problems stemming from classifying filler identifications as either correct or incorrect across the different procedures, we classified responses into those who identified the suspect (guilty or innocent) and those who made another decision (filler, rejection, indeterminate). We conducted two separate hierarchical log-linear analyses (HILOG) with the decision to select the suspect (or not) as the dependent variable to determine whether the rate of suspect selections produced by each of the classifications algorithms varied by procedure, age, and target presence. Odds ratios (OR) are provided as a measure of effect size. ORs are the ratio of event occurrences (e.g., correct responses) to non-event occurrences (e.g., incorrect responses) and are calculated by dividing the odds of an event in one group by the odds in another group. An OR of 1 suggests that the two groups are not different. The interpretation of an OR depends on which event occurrence is used

as the numerator and denominator. For this study, the larger odds were used as the numerator to allow for intuitive interpretation. For example, an OR of 2.00 can be taken to mean that the odds of an event for one group (e.g., correct response) are two times greater than for the other group.

C3 algorithm. The HILOG revealed no significant 4-way, $\chi^2(1) = .07$, $p = .80$, or 3-way interactions, $\chi^2(4) = 2.78$, $p = .60$. The highest level interaction was between two variables, $\chi^2(2) = 101.83$, $p < .001$. Partial associations revealed a 2-way interaction between target presence and suspect identifications, $\chi^2\ (1) = 97.24$, $p < .001$, indicating that the number of suspect identifications was higher in the target-present condition than in the target-absent condition, $z = 10.51$, $p < .001$, *OR* = 7.85, 95% CI [4.98, 12.37]. No significant interactions involving procedure (confidence/categorical) were detected.

Hierarchical (H) algorithm. The HILOG revealed no significant 4-way, $\chi^2(1) = .09$, $p = .77$, or three-way interactions, $\chi^2(4) = 4.04$, $p = .40$. The highest level interaction was between two variables, $\chi^2(2) = 96.63$, $p < .001$. Partial associations revealed a 2-way interaction between target presence and suspect identifications, $\chi^2\ (1) = 94.36$, $p < .001$, indicating that the number of suspect identifications made was higher in the target-present condition than in the target-absent condition, $z = 10.41$, $p < .001$, *OR* = 7.17, 95% CI [4.65, 11.07]. No significant interactions involving procedure (confidence/categorical) were detected.

Table 2

Classification proportions (standard errors) by target presence for confidence and categorical procedures

	Categorical		Confidence (C3)		Confidence (H)	
	TP	TA	TP	TA	TP	TA
Younger						

Suspect	0.41 (0.07)	0.12 (0.04)	0.45 (0.07)	0.10 (0.04)	0.52 (0.07)	0.16 (0.05)
Filler	0.25 (0.06)	0.29 (0.06)	0.10 (0.04)	0.33 (0.06)	-	-
Reject	0.34 (0.06)	0.59 (0.06)	0.45 (0.07)	0.57 (0.06)	0.48 (0.07)	0.78 (0.06)
Indeterminate					0.00 (0.00)	0.07 (0.03)
N	56	59	58	58	58*	58*
Older						
Suspect	0.61 (0.06)	0.14 (0.04)	0.51 (0.06)	0.09 (0.04)	0.54 (0.06)	0.12 (0.04)
Filler	0.15 (0.04)	0.34 (0.06)	0.15 (0.04)	0.24 (0.05)	-	-
Reject	0.24 (0.05)	0.51 (0.06)	0.34 (0.06)	0.67 (0.06)	0.46 (0.06)	0.88 (0.04)
Indeterminate					0.00 (0.00)	0.00 (0.00)
N	75	70	71	66	71*	66*

Note. TP = Target-present lineups; TA = Target-absent lineups. As outlined by Sauer et al., (2008), responses were not classified as filler identifications in the hierarchical (H) confidence method because they were assumed to be known errors and, as such, were classified as potential negative identifications (i.e., rejections). Indeterminate responses are only found in the H confidence method and represent confidence judgments considered too weak to be classified as either identifications or rejections. *The classifications for C3 and H were derived by applying different algorithms to the same confidence ratings (they are not from independent samples).

Accuracy of Positive and Negative Responding

In a forensic setting, investigators would not know whether a suspect is guilty or innocent. Therefore, to further explore the accuracy of the responses from the confidence procedure (cf. the categorical procedure), we divided responses into positive and negative classifications (see Table 3)—that is, those who ‘chose’ from the lineup or cases in which the algorithm returned a positive classification (positive) and those who did not ‘choose’ or cases in which the algorithm returned a negative classification (negative). Indeterminate responses were

neither classified as negative nor positive. It is important to note that any cases of multiple max ratings were classified as rejections (see Sauer et al., 2008), which potentially inflates the rate of negative classifications. For example, many younger children (n = 38, 33%) in the confidence procedure gave a maximum confidence rating higher than zero to multiple lineup members—of those, a third (n = 14, 36%) provided the guilty suspect with one of the maximum ratings. Likewise, many older children (n = 47, 33%) in the confidence procedure gave a maximum confidence rating to multiple lineup members—of those, a quarter (n = 12, 26%) provided the guilty suspect with one of the maximum ratings. Thus, our analyses represent a conservative test of the confidence procedure

The C3 algorithm. A 2 (procedure: confidence-C3, categorical) x 2 (response classification: positive, negative) x 2 (age: 6-8, 9-13) x 2 (accuracy: correct, incorrect) HILOG revealed no three-way interaction between age, procedure, and accuracy, χ^2 (1) = 0.02, p = .89 and no three-way interaction between age, procedure, and response classification, χ^2 (1) = 1.06, p = .30. Thus, we did not find any of the key effects of procedure, or interactions between procedure and age group. The only significant association was between response classification and accuracy, χ^2 (1) = 13.78, p < .001, such that accuracy was higher for negative (.63, SE = 0.06) than positive classifications (.47, SE = 0.17), z = 3.75, p < .01, OR = 1.95, 95% CI [1.37, 2.77].

The hierarchical (H) algorithm. A separate HILOG compared performance using the hierarchical algorithm and categorical procedure, with an adjustment to the responses in the latter to facilitate a fair comparison between conditions. More specifically, filler identifications in the categorical condition were changed from positive classifications into correct (TA) or incorrect (TP) negative classifications in order to mimic the hierarchical algorithm's classification method.

In line with previous research (Sauer et al., 2008), indeterminate responses calculated using the hierarchical (H) classification method were included in the denominator when calculating classification accuracy for the confidence procedure. This HILOG revealed a significant two-way interaction between response classification and accuracy, χ^2 (1) = 14.52, $p < .001$, such that positive responses were more accurate (0.80, *SE* = 0.07) than negative responses (0.64, *SE* = 0.04), $z = 4.10$, $p < .001$, *OR* = 2.30, 95% CI [1.49, 3.56]. No other interactions were found.

Table 3

Accuracy rates, error estimates, and condition sample size for positive and negative responses.

			Positive	Negative	Total
Younger	Categorical	Accuracy	0.38 (0.06)	0.67 (0.05)	0.53 (0.05)
		n	61	54	115
	Confidence (C3)	Accuracy	0.46 (0.06)	0.54 (0.05)	0.50 (0.05)
		n	57	59	116
	Categorical (Adjusted)	Accuracy	0.77 (0.08)	0.61 (0.05)	0.65 (0.04)
		n	30	85	115
	Confidence (H)	Accuracy	0.77(0.06)	0.62(0.04)	0.67 (0.04)
		n[1]	39	73	112
Older	Categorical	Accuracy	0.51 (0.04)	0.67 (0.05)	0.57 (0.04)
		n	92	54	146
	Confidence (C3)	Accuracy	0.52 (0.05)	0.65 (0.04)	0.58 (0.04)
		n	69	68	137
	Categorical (Adjusted)	Accuracy	0.82 (0.05)	0.68 (0.04)	0.73 (0.04)
		n	38	58	146
	Confidence (H)	Accuracy	0.83 (0.05)	0.64(0.03)	0.70 (0.03)
		n	46	91	137

[1] Note that these counts exclude indeterminate responses (younger children *n* = 4; older children *n* = 0) as indeterminate responses cannot be classified by response type. Standard errors are in parentheses and were derived from bootstrapping original data.

How do the Confidence and Categorical Procedures Compare in Diagnosticity and Discriminability?

Diagnosticity. Diagnosticity ratios were used to measure the likelihood that an identified suspect was guilty (i.e., only suspect identification rates were considered). A diagnosticity ratio of 1.0 indicates that the two events (i.e., identified suspect is guilty versus identified suspect is innocent) are equally likely. Departure from 1.0 indicates differences in the probability of these two events. For example, a ratio of 2.0 indicates that children were twice as likely to identify the culprit as the innocent suspect. Thus, procedures with confidence intervals that do not overlap with 1 can be considered diagnostic. The diagnosticity ratios in Table 4 show that when a suspect was identified via the confidence procedure, the suspect identification was at least three times as likely to be guilty than to be innocent regardless of the algorithm or the age group. This result indicates that confidence ratings can be used with child eyewitnesses to diagnose whether an identified suspect is guilty or innocent.

We compared diagnosticity ratios across procedures by calculating .05 inferential confidence intervals (ICI; Tryon, 2001) using a bootstrapping procedure (Palmer et al., 2010). In following the procedure laid out by Palmer et al. (2010), we first resampled the observed data 3000 times and computed a diagnosticity ratio for each of the replicated samples. Then, to more closely approximate a normal distribution, we transformed the diagnosticity ratios to log scale (ln). Finally, the distribution was used to estimate the variance needed to calculate inferential confidence intervals (see Tryon, 2001). The inferential confidence intervals in Table 4 have been converted from log scale back to their original unit. Table 4 shows diagnosticity ratios were similar for the confidence (C3 and H) and categorical procedures.

Table 4

Diagnosticity ratios and .05 inferential confidence intervals (ICI)

			.05 ICI [Lower Limit, Upper Limit]	
		Diagnosticity ratio	Confidence (C3)	Confidence (H)
Younger Children	Categorical	3.46	[1.84, 6.50]	[1.84, 6.48]
	Confidence (C3)	4.33	[2.26, 8.31]	-
	Confidence (H)	3.33	-	[1.97, 5.62]
Older Children	Categorical	4.35	[2.71, 6.99]	[2.72, 6.96]
	Confidence (C3)	5.58	[2.87, 10.80]	-
	Confidence (H)	4.42	-	[2.48, 7.85]

Note: .05 ICI = Inferential Confidence Intervals. ICIs allow for pairwise comparisons between diagnosticity ratios for each confidence procedure (i.e., C3 or H) and the categorical procedure within each age group. That is, comparisons of ICIs should be made within each column and not across age groups.

Discriminability. Although diagnosticity ratios have traditionally been used to provide an overall assessment of lineup performance, this method has recently been criticized for its susceptibility to influences of response criterion (Wixted & Mickes, 2012). To avoid problems of inflated diagnosticity ratios, Mickes, Moreland, Clark and Wixted (2014) argued for discriminability in eyewitness identification tasks to be measured using the signal-detection statistic, *d'*, in the absence of the ability to calculate receiver operator characteristic (ROC) curves. However, as Palmer, Brewer and Weber (2010) point out, eyewitness identification decisions are not simple binary decisions (hit or miss for target-present stimulus; false alarm or

correct reject for target-absent stimulus), but in fact include an additional response type: filler selections. Using traditional SDT methods to calculate *d'*, these filler identification classifications pose a problem because they can be classified as either a false alarm or a miss, but are not always genuine examples of either response type. Some researchers have opted to treat filler identifications in target-present lineups as misses, thus excluding the responses in the calculation of *d'* (e.g., Meissner, Tredoux, Parker, & MacLin, 2005; Mickes et al., 2014). In doing so, *d'* can be conceptualized as an index of how well a group discriminates between guilty and innocent suspects, with a higher number indicative of better discrimination. Using this method, the *d'* values for the C3, hierarchical, and categorical procedures for younger children were 1.13, 1.06, and 0.96, respectively, and for older children were 1.35, 1.26, and 1.36, respectively. However, a witness who incorrectly selects a filler is not, in memory or decision-making terms, equivalent to a witness who views a lineup and, for whatever reason, elects not to pick anyone.

Eyewitness identification can be conceptualized as a compound decision that involves two tasks: (a) *detection* or determining if the target is present in a group; and (b) *identification* or determining the correct target from the group (Duncan, 2006; Macmillian & Creelman, 1991; Palmer et al., 2010). To account for the complexity of compound decisions, Duncan (2006) proposed a compound decision model of signal detection theory (SDT-CD). SDT-CD generates expected probabilities of detection and identification and can be applied to estimate discrimination and response bias for lineup identification tasks (e.g., Palmer et al., 2010).

For the detection component, the model generates estimates of how often a decision maker will choose in a target present lineup (also referred to as the hit rate in this model) and how often a decision maker will choose in a target-absent lineup (also referred to as the false

alarm rate in this model). In the SDT-CD model, a positive response to the detection component is dependent upon one of two decision rules that may be used by a decision maker. First, the *independent observation rule* proposes that each stimulus in an array is assessed separately against a single criterion. A positive response is made when at least one stimulus in the array meets the criterion. The second, or more global rule, is the *integration rule* in which each stimulus in the array is assigned a value and the sum of these values are compared against a criterion to make a decision. A positive response is made when the sum of the stimulus values meets or surpasses the criterion.

The SDT-CD model is designed to generate expected probabilities of the identification component—or how often a decision maker will choose the correct target in a target-present lineup (correct identification rate). The decision rule for the identification component assumes that the decision maker will choose from the target-present array based on the probability that the similarity of the chosen stimulus with the intended target exceeds all of the remaining stimuli.

Using these expected probabilities, the model then generates estimates of discriminability (*d'*) and response bias (*c*). These three expected probabilities (hit, false alarm, and correct identification rates) are then compared to the observed data. If good fit is found, then the model-generated estimates of *d'* and *c* can be used as reasonable estimates of discriminability and response bias of the observed data (Duncan, 2006; Palmer et al., 2010). Under the SDT-CD model, higher positive values of *d'* indicate a respondent's ability to distinguish between the target and non-target stimuli in the identification and detection components, with a *d'* of zero indicating no discrimination. A *c* statistic indicates willingness to choose a target from the stimulus array. Positive *c* values indicate conservative responding, negative values indicate lenient responding, and a value of zero indicates no bias in responding.

We used the SDT-CD model to compare discriminability and response bias across the two lineup procedures. The SDT-CD model was designed to include filler identifications in the calculations. As the hierarchical confidence classification method does not provide this category, only the C3 confidence method was assessed. Following Palmer and colleagues' (2010) approach, the best-fitting combination of *d'* and *c* statistics were identified by comparing observed and model-generated response probabilities using likelihood ratio G-statistics (Sokal & Rohlf, 1981). During the model selection process, all parameters were considered within the range noted by Palmer and colleagues (2010; -1.59 to 4.01 for *d'* and -3 to 3 for *c*). Previous eyewitness research using SDT-CD has used both simultaneous (e.g., Palmer et al., 2010) and sequential presentation methods (Palmer & Brewer, 2012). The integration rule has been previously applied to sequential lineup data (Palmer & Brewer, 2012). However, the proposed computation of a total score upon which to make a decision in the integration rule is conceptually problematic for use with a sequential procedure as well as with the C3 algorithm. As a result, only the independent observation model was considered.

For each condition, G-tests were conducted to compare observed and model-generated expected frequencies for all response types (see Table 5). Using *Excel* solver, the total G-statistics for each of the tests were optimized to find the best-fitting combination of *d'* and *c* estimates (see Table 6). The results showed that the model fit the data for the independent rule, with both procedures producing non-significant total G values (all total Gs ($df = 3$) < 4.63, $p > .10$). The one exception to finding good fit was with young children's C3 response classifications ($G = 11.13$, $p < .001$). This indicates that the independent observation model does not provide a good fit for younger children's responses to the confidence procedure and, thus, interpretations of *d'* and *c* should be made with caution. As good fit between our data and the model data was

found for the remaining groups, the model-generated estimates of *d'* and *c* were used to represent discriminability and response bias (Duncan, 2006). Table 5 provides a breakdown of the SDT-CD model results.

Table 5

Observed and (best-fitting) Independent Observation (IO) model-generated response proportions for target-present and target-absent lineups.

		Target-Present			Target-Absent	
		Correct ID	Filler ID	Rejection	Filler ID	Correct Rejection
Younger	**Categorical**					
	Observed	0.41	0.25	0.34	0.41	0.59
	IO Model	0.40	0.28	0.31	0.38	0.62
	Confidence(C3)					
	Observed	0.45	0.10	0.45	0.43	0.57
	IO Model	0.42	0.25	0.34	0.33	0.67
Older	**Categorical**					
	Observed	0.61	0.15	0.24	0.48	0.52
	IO Model	0.60	0.22	0.18	0.42	0.58
	Confidence(C3)					
	Observed	0.51	0.15	0.34	0.33	0.67
	IO Model	0.50	0.21	0.30	0.29	0.71

Note. ID = Identification

Next, a bootstrapping procedure was used to estimate variance for the *d'* and *c* statistics. In line with previous research (Palmer et al., 2010; Weber & Brewer, 2006), these variance

estimates were used to create .05 inferential confidence intervals for the statistics. Specifically, the response frequencies for each condition (observed data) were used as a sampling distribution from which 3000 replication data sets were randomly drawn (see Palmer et al., 2010). Next, optimized *d'* and *c* statistics were calculated for each of these 3000 data sets that provided the distribution needed to calculate inferential confidence intervals (see Tryon, 2001). Non-overlapping confidence intervals are indicative of a significant difference. As seen in Table 6, there are minimal differences in the estimated *d'* values between the confidence (C3) and categorical conditions. In addition, the confidence procedure is associated with significantly more conservative responding than the categorical procedure. One consideration is that the SDT-CD model assumes that the suspect is selected from the same distribution as the fillers. Given that the innocent suspect in the present study was selected as the lineup member who most closely resembled the target, this assumption was violated. Although this may have affected the estimation of *d',* we can think of no reason the violation would differentially affect estimation across conditions. Nevertheless, caution should be taken when interpreting these values. Further, SDT indices should be used for relative comparisons, not as an absolute index of discrimination or bias.

Table 6

Independent Observation (IO) model-generated SDT-CD estimates of discriminability (*d'*), response bias (*c*)

	Condition	Estimated discriminability			Estimated response bias		
			.05 ICI			.05 ICI	
		d'	Lower	Upper	*c*	Lower	Upper
Younger	Categorical	1.61 (0.20)	1.33	1.88	0.76 (0.10)	0.62	0.90

	Confidence (C3)	1.72 (0.19)	1.45	1.99	0.80 (0.10)	0.66	0.94
Older	Categorical	2.06 (0.16)	1.83	2.29	0.48 (0.09)	0.35	0.61
	Confidence (C3)	1.98 (0..17)	1.74	2.21	0.74 (0.09)	0.62	0.86

Note: Estimated standard errors are in parentheses; .05 ICI = Inferential Confidence Intervals. ICI allow for comparisons of *d'* or *c* across procedures (C3 with Categorical) within each age group.

Profiles of Individual Confidence-Accuracy Relationships

The results presented thus far indicate the confidence procedure and the categorical procedure produce comparable response accuracy. This provides evidence that, at a group level, we can use children's confidence ratings to infer lineup responses. However, in applied settings, criminal investigators will be interested in the accuracy of individual witness responses—not an entire group. To help with this, we used Brewer and colleagues' (2012) profile analysis to highlight what individual sets or patterns of confidence ratings were likely (versus not likely) to indicate accurate discrimination between previously seen and previously unseen faces.

Brewer and colleagues used a discrepancy score between the maximum rating and the next-highest rating on a 100-point scale. However, because we used a 5-point scale when adapting the task for children, variability in responses is smaller than for Brewer and colleagues. To overcome this, the profile analysis provides a classification-accuracy rate as a function of discrepancy between the maximum and the average of all other confidence values. For example, if the maximum confidence was 100% (4) and the average of the remaining ratings was 25% (1), the discrepancy would be 75% (3). We converted the 5-point scale into a grouped-discrepancy score (0 = 0%, 1 = 25%, 2 = 50%, 3 = 75%, 4 = 100%).

We then completed the profile analysis using the same stipulations outlined by Brewer et al. (2012). That is, we only examined lineups in which a single maximum confidence rating was made and excluded filler identification responses in target-present lineups on the basis that they

would be a known incorrect selection in applied settings. Results of Brewer et al.'s (2012) profile analysis provide information about the probability that a suspect was the target at each level of discrepancy. Note that we also conducted the profile analyses using the same method described by Brewer and colleagues (2012) and this information can be found in the Online Supplementary Materials.

As seen in Table 7, the results of the profile analysis show a similar linear relationship between discrepancy and accuracy that has been observed with adults (Brewer et al., 2012). One exception was that we did not observe 100% accuracy at the 100% level of discrepancy. For both the younger and older children, one child in a target absent condition incorrectly reported high confidence that an innocent lineup member was the guilty suspect. Thus, this procedure did not fully negate the problems associated with children's identifications (e.g., lack of task understanding, pressure to assign a high value to indicate a 'choice').

Table 7

Proportion of correct decisions and number of responses for each category of discrepancy (between the maximum and the average of all other confidence values) for children in the confidence condition.

	Profile Analysis			
	Younger Children		Older Children	
Discrepancy	Number of responses	Proportion correct	Number of responses	Proportion correct
100	5	0.80	7	0.86
≥75	22	0.68	24	0.88
≥50	56	0.46	47	0.72

≥25	83	0.36	73	0.53
>0	87	0.34	73	0.53

Note. Confidence was rated on a scale of 0 to 4 and was converted to a 0% to 100% scale.

Discussion

A frequently cited problem with child witnesses is that they are too lenient in their decisions—that is, they choose too frequently with categorical procedures (see Fitzgerald & Price, 2015). This high choosing rate may be partly explained by their use of overly-lenient response criteria (Dunlevy & Cherryman, 2013). We hypothesized that using this confidence procedure, and wresting control of the decision criterion away from the witness, would reduce problems associated with use of overly-lenient decision criteria and, in turn, reduce inappropriate choosing behavior. To explore this hypothesis, we first examined whether children could appropriately use confidence ratings to indicate the degree of match between previously seen and unseen faces. Then, we collapsed children's ratings into categorical responses (using classification algorithms) to compare performance with a sequential procedure.

Utility of the Confidence Procedure

First, we assessed whether or not children could use the confidence rating procedure to accurately discriminate between previously seen and unseen faces. This research provides early evidence that confidence ratings can provide meaningful information about children's recognition memory. This conclusion is based on three analyses. First, ANDI scores demonstrated that both younger (.20) and older (.24) children were able to use confidence ratings to discriminate between previously seen and unseen faces. Second, the algorithms were able to classify children's responses such that suspect identification accuracy was above chance (50%). Third, the observed linear pattern between discrepancy and classification accuracy rates in the profile analysis (Table 7) demonstrates that children's confidence ratings can be used to

effectively discriminate guilty from innocent suspects. These data demonstrate that both age groups of children can use confidence ratings to index likely guilt in a way that reduces or mitigates decision criteria influences, and permits a probabilistic assessment of identification evidence. This crucial finding provides the foundation for further exploration of procedures based on children's confidence assessments.

Confidence Ratings versus Categorical Identifications

We also compared the responses from the confidence paradigm with those produced from a categorical procedure. A handful of studies have examined the confidence rating procedure with adult witnesses (e.g., Brewer et al., 2012; Sauer et al., 2008, 2012; Weber & Varga, 2012). Despite differences in the methodologies used by these studies and the present study (e.g., presentation style, age, confidence rating scale), a generally consistent finding is that classification algorithms can be applied to confidence ratings to produce results that are at least comparable, and often superior to performance using categorical procedures. In our research, the C3 and hierarchical confidence classification method produced very similar results to the categorical procedure for both age groups. Specifically, both classification algorithms produced comparable suspect identifications rates as well as comparable positive and negative response accuracy rates. The comparable performance provides evidence that we can use confidence ratings provided by children to index recognition memory at a level that is equivalent with a traditional categorical lineup procedure. These results support previous findings that children possess the metacognitive abilities to report confidence ratings that are sensitive to the nature of the stimuli (i.e., accurately discriminate between old and new stimuli; Hiller & Weber, 2013).

Age. Although exploring age differences was not a central focus of the present study, future research into age differences may help to better understand the mechanisms at play. We

did not observe an age-related effect on overall performance. This is somewhat surprising given that metacognitive development has been found to be an obstacle facing young children when providing confidence ratings (e.g., Brewer & Day, 2005; Keast et al., 2007). A lack of age effects may be explained by the coarse dichotomization of the age variable in the present study. Alternatively, our results could suggest that changing the decision task (from a task requiring children to compress all the information from a lineup into a single decision to one that focuses on each person in the lineup) to mitigate effects of criterion placement may attenuate age differences (Hiller & Weber, 2013).

Future Research Considerations

Our understanding of the value that the confidence procedure holds for use with child witnesses is in its infancy. This is the first study, to our knowledge, to apply this paradigm to a child sample. When we turn to the adult literature, there are some inconsistencies in the findings regarding the impact of using a confidence paradigm relative to a yes-no paradigm. These inconsistencies are due to the exploratory nature of the confidence paradigm and the classification algorithms involved. The classification methods that were first introduced (see Sauer et al., 2008) have evolved over time (Weber & Varga, 2012) and, we expect, will continue to do so.

For these reasons, this research is currently most informative from a cognitive perspective, as it is premature to apply the confidence procedure to legal settings. However, there is value in considering the impact this sort of procedure may have on the legal system. For example, how will legal decision makers consider evidence based on confidence ratings, rather than a clear, categorical decision? . As indicated by previous research, hearing an eyewitness state ‘that’s the guy I saw’ is a powerful and persuasive form of evidence (e.g., Boyce, Lindsay,

& Brimacombe, 2008; Cutler, Penrod & Dexter, 1990). Not providing that information to decision makers in a legal setting may prove to be a challenge to those expecting finality in a witness statement. However, when considering the purpose of conducting a lineup task, there is a clear space for use of a confidence procedure in the legal system. And, although less traditional, Sauer, Palmer, and Brewer (2017) recently reported that mock-jurors are receptive to non-categorical forms of identification evidence and, with coaching, can appropriately evaluate this type of evidence. As Charman and Wells (2007) point out, the aim of a police lineup is not to test the eyewitness but, rather, to gather evidence as to the guilt of a possible suspect. From this perspective, the confidence procedure may provide more valuable eyewitness evidence than the current lineup paradigms available to investigators.

Confidence rating-based identification evidence has several advantages over a categorical identification. For instance, confidence ratings for each lineup member provide investigators with multiple points of information, including which member best matches a child's memory of a perpetrator as well as the degree to which the best match is preferred, relative to the other members. Importantly, although collapsing patterns of confidence ratings into categorical classifications is useful for comparing performance against a traditional lineup procedure, this actually obscures some of this useful information. Recognition memory is not an "all or nothing" construct: The strength of recognition falls on a continuum. Thus, we argue that there is merit in encouraging legal decision-makers to shift from interpreting identification evidence as a clear-cut indication of guilt toward a more probabilistic treatment of the evidence (Sauer & Brewer, 2015). Moving from a categorical treatment of identification evidence to a ratings-based approach recognizes this distinction. The ratings-based approach allows for graded evidence against a suspect based on both the strength of the witness's recognition of the suspect and the

witness's ability to discriminate between the suspect and other lineup members. The potential value of this approach is evident in the linear relationship observed in the profile analysis reported in Table 7. As the level of discrepancy increases so, too, does the likely guilt of the suspect (see also Brewer et al., 2012). Thus, the most important aspect of the current findings may not be the actual accuracy rates observed, but the evidence that even younger children can use confidence ratings to discriminate guilty from innocent suspects.

Moreover, many children provided multiple maximum ratings. In keeping with previous research (Sauer et al., 2008), responses from those who provided multiple max ratings were classified as rejections. However, there are nuances in these multiple maximum responses that may provide valuable information about memory strength. For example, does providing a maximum rating to four faces indicate a weaker memory than providing a maximum rating to only two faces? How informative is a child's memory when he or she provides a maximum rating to the suspect, along with one other lineup member (versus two or three others)? There is a need to further explore the value of the confidence procedure as probabilistic evidence of suspect guilt, including whether the number of maximum ratings provided (and who they are given to) can be used as a supplemental index of recognition memory.

The need for independent replication and applying this procedure more broadly are natural next steps. For example, the ecological validity of stimulus materials (e.g., live and/or emotionally arousing events) and lineup presentation methods (e.g., simultaneous presentation; video or live lineups) should be considered in future research. There is also a need to compare classification of confidence ratings to a sequential procedure that does not contain any interim confidence ratings. While we opted to include confidence ratings following each categorical decision to examine whether the confidence ratings can be used in conjunction with a categorical

decision, it is worthwhile to directly compare confidence ratings to a more 'pure' categorical rating of the overall lineup decision.

Limitations. Eliciting a confidence rating for each lineup member may not completely avoid decision criteria influences that are observed with a categorical decision. Just as children may feel pressure to choose from a traditional lineup task, they may also feel pressure to provide at least one high rating. Children did not appear to use the confidence scale in a dichotomous or binary way (i.e., primarily selecting from either end of the scale); however, some children may have still felt pressure to provide a high rating to at least one lineup member. Because we did not attempt to asses if children felt obligated to assign a maximum value, we do not know the extent to which the confidence procedure assisted in avoiding decision criteria influence.

Finally, given that this was an initial exploration of children's use of confidence ratings and we did not focus on exploring developmental differences, we did not have a sample size large enough to capture the nuanced differences that can be expected for children aged 6-7, from those who are 8-9, and beyond. Therefore, the lack of observable differences between age groups may be due to exploring age categorically, rather than continuously. Going forward, it would be beneficial to focus on a narrower age range of children or explore age continuously in order to learn more about developmental differences in use of confidence ratings.

Conclusions

These findings provide evidence that confidence ratings are a useful index of recognition for child eyewitnesses. When applied to child eyewitnesses, the confidence procedure can be used to provide categorical assessments of guilt that work as well as a standard, sequential procedure. However, the more important implication is that children can use a ratings-based procedure to discriminate between previously seen and unseen faces. In turn, this invites more

nuanced and empirical consideration of ratings-based identification evidence as a probabilistic index of guilt that may attenuate problematic social influences on child witnesses' decision criteria. Taken together, the present findings suggest that further refining of the procedure, especially the use of children's confidence ratings as probabilistic evidence of suspect guilt, is well worth consideration.

References

Beal, C. R., Schmitt, K. L., & Dekle, D. J. (1995). Eyewitness identification of children: Effects of absolute judgments, nonverbal response options, and event encoding. *Law and Human Behavior*, *19*, 197-216. doi: 10.1007/BF01499325

Brewer, N., & Day, K. (2005). The confidence-accuracy and decision latency-accuracy relationships in children's eyewitness identification. *Psychiatry, Psychology and Law*, *12*, 119-128. doi: 10.1375/pplt.2005.12.1.119

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indices of memory processes not just markers of accuracy. *Applied Cognitive Psychology, 22,* 827-840

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23,* 1209-1214. doi:10.1177/0956797612441217

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

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science, 20*, 24-27.

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: Evidence from observational analysis of young children's behavior during problem-solving. *Metacognition and Learning*, *7*, 197-217. doi: 10.1007/s11409-012-9091-2

Charman, S. D., & Wells, G. L. (2007). Applied lineup theory. In R. C. L. Lindsay, D. Ross, D. Read, & M. Toglia (Eds.), *Handbook of Eyewitness Psychology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence. *Law and Human Behavior, 4,* 243-260.

Duncan, M. J. (2006). A signal detection model of compound decision tasks (Tech. Rep. No. TR2006-256). Toronto, ON; Defence Research and Development Canada.

Dunlevy, J. R., & Cherryman, J. (2013). Target-absent eyewitness identification line-ups: Why do children like to choose. *Psychiatry, Psychology and Law, 20*, 284-293. doi:10.1080/13218719.2012.671584

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Fitzgerald, R. J., & Price, H. L. (2015). Eyewitness identification across the lifespan: A meta-analysis of age differences. *Psychological Bulletin, 141,* 1228-1265. doi: 10.1037/bul0000013

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R.V. Kail, Jr., and J.W. Hagen (Eds.), *Perspectives on the Development of Memory and Cognition,* Hillsdale, NJ, Erlbaum.

Haller, E. P., Child, D. A., & Walberg, H. J. (1988). Can comprehension be taught? A quantitative synthesis of "metacognitive" studies. *Educational Researcher, 17,* 5-8.

Hiller, R. M., & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2,* 185-191. doi: 10.1016/j.jarmac.2013.07.001

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1304-1316.

Keast, A., Brewer, N., & Wells, G.L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology, 97,* 286-314. doi: 10.1016/j.jecp.2007.01.007

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,*490–517. doi: 10.1037/0033-295X.103.3.490

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*, 405-437. doi: 10.1006/jecp.2000.2612

Leippe, M. R. (1980). Effects of integrative memorial and cognitive processes on the correspondence of eyewitness accuracy and confidence. *Law and Human Behavior, 4*, 261-274. doi: 10.1007/BF01040618

Lindsay, R. C., Pozzulo, J. D., Craig, W., Lee, K., & Corber, S. (1997). Simultaneous lineups, sequential lineups, and showups: Eyewitness identification decisions of adults and children. *Law and Human Behavior*, *21*, 391-404. doi: 10.1023/A:1024807202926

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute *d'*, not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition, 3*, 58-62. doi: 10.1016/j.jarmac.2014.04.007

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behavior, 16,* 575–593. doi:10.1007/BF01044624

Norman, D. A., & Wickelgren, W. A. (1965). Short-term recognition memory for single digits and pairs of digits. *Journal of Experimental Psychology, 70,* 479-489.

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36,* 247-255. doi: 10.1037/h0093923

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied, 16,* 387-398. doi: 10.1037/a0021034.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*(1), 55-71. doi: 10.1037/a0031602

Parker, J. F., & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adults' eyewitness identifications. *Law and Human Behavior, 17,* 11-26. doi: 10.1007/BF01044534

Perfect, T. J., & Weber, N. (2012). How should witnesses regulate the accuracy of their identification decisions: One step forward, two steps back? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1810-1818. doi: 10.1037/a0028461

Police Executive Research Forum. (2013). *A national survey of eyewitness identification processes in law enforcement agencies.* Washington, DC: Police Executive Research

Forum. Retrieved from http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf

Pozzulo, J. D., Dempsey, J., Bruer, K., & Sheahan, C. (2012). The culprit in target-absent lineups: Understanding young children's false positive responding. *Journal of Police and Criminal Psychology, 27, 55-62*. doi: 10.1007/s11896-011-9089-8.

Pozzulo, J. D., & Lindsay, R. C. L. (1998). Identification accuracy of children versus adults: A meta-analysis. *Law and Human Behavior, 22,* 549-570. doi: 10.1023/A:1025739514042

Pozzulo, J. D., & Lindsay, R. C. L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. *Journal of Applied Psychology, 84,* 167-176. doi: 10.1037/0021-9010.84.2.167

Price, H. L., & Fitzgerald, R. J. (2016). Face-off: A new identification procedure for child eyewitnesses. *Journal of Experimental Psychology: Applied, 22*, 366-380. doi: 10.1037/xap0000091

Roebers, C. M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38,* 1052–1067. doi: 10.1037/0012-1649.38.6.1052

Roebers, C. M., & Howie, P. (2003). Confidence judgments in event recall: Developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85*, 352–371. doi: 10.1016/S0022-0965(03)00076-6

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Chichester: Wiley Blackwell.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*, 528-547. doi: 10.1037/a0012712

Sauer, J. D., Brewer, N., & Weber, N. (2012). Using confidence ratings to identify a target among foils. *Journal of Applied Research in Memory and Cognition, 1*, 80-88. doi: 10.1016/j.jarmac.2012.03.003

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347. doi: 10.1007/s10979-009-9192-x

Sauer, J. D., Palmer, M. A., & Brewer, N. (2017). Mock-Juror Evaluations of Traditional and Ratings-Based Eyewitness Identification Evidence. *Law and Human Behavior.* Advance online publication. http://dx.doi.org/10.1037/lhb0000235

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. Dunlosky & R. A. Bjork (Eds.), A handbook of metamemory and memory. Mahwah, NJ: Erlbaum.

Sokal, R. R., & Rohlf, F. J. (1981). *Biometry* (2nd ed.). San Francisco: Freeman.

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

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*, 99. doi: 10.1037/a0021650

Trow, W. C. (1923). The psychology of confidence. *Archives of Psychology, 67*, 3–48.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371-386. doi: 10.1037/a0013158

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 582–600. doi: 10.1037//0278-7393.26.3.582

Vickers, D. (1979). *Decision processes in visual perception.* New York: Academic Press.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgements. *Journal of Experimental Psychology: Applied, 10,* 156-172. doi: 10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy, and response latency. *Applied Cognitive Psychology*, *20*(1),17-31. doi: 10.1002/acp.1166

Weber, N., & Perfect, T. (2012). Improving eyewitness identification accuracy by screening out those who say they don't know. *Law and Human Behavior, 36,* 28-36. doi: 10.1007/s10979-011-9269-1

Weber, N., & Varga, M. (2012). Can a modified lineup procedure improve the usefulness of confidence?. *Journal of Applied Research in Memory and Cognition*, *1*, 152-157. doi: 10.1016/j.jarmac.2012.06.007

Wellman, H. M. (1978). Knowledge of the interaction of memory variables: A developmental study of metamemory. *Developmental Psychology*, *14*, 24–29. doi: 10.1037/0012-1649.14.1.24

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83,* 360–376. doi: 10.1037/0021-9010.83.3.360

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7,* 45–75. doi: 10.1111/j.1529-1006.2006.00027.x

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54,* 277-295. doi: 10.1146/annurev.psych.54.101601.145028

Weston, H.E., Boxer, P., & Heatherington, L. (1998). Children's attributions about family arguments: Implications for family therapy. *Family Process, 37,* 35-49. doi: 10.1111/j.1545-5300.1998.00035.x

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316-347. doi: 10.1016/0022-2496(66)90018-6

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7,* 275-278. doi: 10.1177/1745691612442906

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110,* 611-617. doi: 10.1037/0033-2909.110.3.611

Appendix A – Confidence Cup Scale