Applied Cognitive Psychology, Appl. Cognit. Psychol. **31**: 448–466 (2017)
Published online 14 August 2017 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.3342

A Repeated Forced-choice Line-up Procedure Provides Suspect Bias Information with No Cost to Accuracy for Older Children and Adults

KAILA C. BRUER[1]* and HEATHER L. PRICE[2]

[1]*Department of Psychology, University of Regina, Regina, Canada*
[2]*Thompson Rivers University, Kamloops, Canada*

Summary: In two experiments and one follow-up analysis, we examined the impact of using a repeated forced-choice (RFC) line-up procedure with child and adult eyewitnesses. The RFC procedure divides the identification task into a series of exhaustive binary comparisons that produces not only traditional line-up information (identification decision and confidence) but also information about witness' selection behavior. Experiment 1 revealed that younger children (6- to 8-year-olds) struggled with the RFC procedure, while older children (9- to 11-year-olds) performed as well with the RFC procedure as with a simultaneous procedure (with wildcard). Experiment 2 replicated this comparable performance with adults. Witnesses' suspect selection behavior during the RFC was predictive of identification accuracy for older children and adults. A model examined the additional information provided by the RFC in experiments 1 and 2 and provided evidence that witnesses' patterns of responding can be used to estimate suspect selection bias (a proxy for suspect recognition strength) associated with individual line-up decisions.

'The spectre of erroneous convictions based on honest and convincing, but mistaken, eyewitness identification haunts the criminal law' (Justice David Doherty; R. v. Quercia, 1990). As Justice Doherty notes, even with the best of intentions, eyewitness identifications can be inaccurate. Adults, and particularly children, struggle with line-up identifications (Fitzgerald & Price, 2015). Misidentifications are problematic because they can implicate an innocent suspect and contribute to a wrongful conviction. Because many justice systems rely on eyewitness evidence when determining fact in criminal cases, a great deal of effort has been made to learn as much information as possible about a witness' memory when assessing the likelihood of guilt.

Traditional line-up procedures typically provide one key piece of information: who in the line-up, if any, the witness believes to be the perpetrator. This information is obtained by asking a witness to make a single, categorical decision when presented with a suspect placed amongst fillers (Steblay, Dysart, & Wells, 2011). This decision can often be supported using the witness' subsequent confidence rating. Post-identification confidence ratings can be indicative of accuracy for adults who choose from a line-up when confidence ratings are collected under ideal circumstances, such as no post-identification feedback (e.g., Brewer & Wells, 2006; Sporer et al., 1995). However, children demonstrate greater overconfidence and poorer calibration than adults (Keast, Brewer, & Wells, 2007), although some researchers have found evidence that increases in children's confidence is related to increasing accuracy (Hiller & Weber, 2013). Thus, confidence ratings may be a useful indication of discrimination to support a traditional line-up identification.

Given the high value placed on eyewitness evidence in the justice system (e.g., Beaudry et al., 2015; Bradfield & Wells, 2000), these traditional approaches to administering line-ups have been scrutinized (e.g., Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Dupuis & Lindsay, 2007; Sauer, Brewer, & Weber, 2008). Wells, Memon, and Penrod (2006), for example, argued that eyewitness researchers have been limited by working within boundaries placed on them through the legal system, and thus, there is a need to develop alternative approaches to understanding how to improve eyewitness performance. In response, there have been recent attempts to move away from traditional line-up approaches to collect more information about memory through alternative procedures and/or analyses (e.g., Bayesian modeling, Smith, Lindsay, & Wells, 2016; confidence procedure, Sauer et al., 2008; grain-size line-up procedure, Horry, Brewer, & Weber, 2016; and receiver operating characteristics, Wixted & Mickes, 2014).

In the present research, we continue this exploration of alternative approaches to understanding eyewitness decisions. Specifically, we explored a new line-up procedure—the *repeated forced-choice* (RFC) procedure. The RFC procedure divides the identification task into a series of exhaustive binary comparisons and was designed not only to produce the same key piece of information that traditional line-ups do (i.e., a categorical decision) but also to provide additional memorial information that can be used to understand a witness' line-up decision.

Using a forced-choice model of face recognition

The RFC procedure is an adapted version of the two-alternative forced-choice method, which is one of two frequently used methods to assess face recognition accuracy in the basic memory literature (Jang, Wixted, & Huber, 2009), including with child populations, aged 4–11 years (e.g., Hood, Macrae, Cole-Davies, & Dias, 2003). The other common assessment method is the old/new format that involves serially presenting previously seen faces (multiple targets) amongst a series of previously unseen faces (lures) after which participants indicate whether they have seen the face before (old) or the face is new (new).

In contrast, the two-alternative forced-choice method tests how accurately a participant can discriminate between two

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

visual stimuli that are presented simultaneously (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). During this task, participants see a single target and then during the testing phase, they are shown two faces, one previously seen and one previously unseen, and asked to decide which is the most similar to the target face (e.g., Jenkins, Lavie, & Driver, 2005). The forced-choice recognition paradigm frequently demonstrates more accurate recognition decision-making than the old/new method for both children and adults (e.g., Jang et al., 2009), especially in situations with efficient encoding conditions (good view and warned about memory testing; Deffenbacher, Leu, & Brown, 1981).

The methodology used in facial recognition research is difficult to apply to line-up research. Facial recognition research using an alternative forced-choice method typically presents multiple targets to a participant (e.g., Deffenbacher et al., 1981). In contrast, in a typical eyewitness context, only one target is present (at encoding and at test), who is then placed amongst a series of distracters for recognition. Moreover, a critical factor that likely contributes to an advantage of a forced-choice procedure over old/new judgments is that the observer knows that one of the two items presented is a target. The observer can then simply choose the option that has a stronger match to their memory of the previously seen face. In contrast, in a line-up context, witnesses may be shown a target-absent (TA) photo array.

Forced-choice procedures have recently been introduced in the literature. In 2016, Price and Fitzgerald introduced the *face-off* procedure that involves showing two images[1] and asking a witness to choose which one looks most similar to a perpetrator. The 'loser' is removed from the procedure, and the 'winner' is subsequently compared with another 'winner' image. This procedure continues until the final picture is presented to the witness, who then must decide whether it is a picture of the guilty perpetrator. The face-off procedure evokes repeated viewing of only some of the line-up members, which may create a procedure-induced commitment bias (Lampinen, Neuschatz, & Cling, 2012). One way to maintain the advantage of repeated choosing but mitigate the risk of a commitment effect is by ensuring that all line-up members are presented an equal number of times. Prior to development of the recent face-off procedure (Price & Fitzgerald, 2016), no line-up researcher to our knowledge has utilized a two-alternative forced-choice model to measure eyewitness accuracy.[2]

In the present work, we tested a new line-up procedure that also utilizes a forced-choice assessment approach. However, to remove potential issues stemming from repeatedly seeing only some faces, we introduced an element of repetition that resulted in a repeated, forced-choice procedure. With the RFC procedure, witnesses are shown several pairs of line-up members and are told that, out of all of the pairs they will be shown, only one person may (or may not be) the target person. To adjust for this, instructions were adapted from previous developmental line-up work (e.g., elimination line-up, Pozzulo & Lindsay, 1998; face-off procedure, Price & Fitzgerald, 2016). When shown each pair, witnesses are asked, 'Which of these two looks MOST like the person you saw?' This contrasts with a traditional alternative forced-choice model in which each pair contains one of multiple targets and participants are asked, 'Which of these two faces is the one you saw' Although the task is similar, the focus for participants is quite different in these two task (which is most similar versus which is the one you saw) that may produce rather different results than what has been observed in the facial recognition literature (i.e., superior recognition accuracy with forced choice).

Repeated forced-choice procedure

In the RFC design, witnesses are presented with all possible pairings of line-up members and asked, for each pair, to choose who looks MOST like the target (i.e., 'most-similar' judgments). Once all pairings are decided, witnesses are shown the two line-up members that were selected most often and asked to make another 'most-similar' decision between them. Lastly, the final remaining line-up member is compared against a blank silhouette picture (a *wildcard* indicating the absence of the target; Zajac & Karageorge, 2009), and witnesses decide if the remaining picture is, in fact, the target. There are two key features of the RFC procedure that were expected to impact line-up performance: reduced stimulus set size (i.e., showing only two pictures at once) and repetition.

Stimulus set size

Simultaneous procedures require witnesses to compress a substantial amount of complex information into a single decision. For example, when presented with a simultaneous line-up, a witness is required to compare all faces with their memory of a target, evaluate the strength of those comparisons, and make a decision. In visual search tasks, such as a line-up, attention attributed to the task is impacted by the set size (Palmer, 1994). Presenting witnesses, especially children, with a task that requires visual assessment of multiple options likely depletes their cognitive resources (Alvarez & Cavanagh, 2004; Pozzulo & Lindsay, 1999).

In a recent study introducing the face-off line-up procedure with school-aged children (aged 6 to 15 years old), Price and Fitzgerald (2016) found that reducing the stimulus set size to a series of pairs improved the accuracy of children's responding in TA line-ups, relative to the simultaneous procedure. However, there was some evidence that this TA benefit may come at a cost to correct choosing from the target-present (TP) line-ups. Price and Fitzgerald argued that showing two faces at once can be advantageous because it provides children with a clear structure to guide their use of relative judgment (i.e., comparison across multiple faces; Lindsay & Wells, 1985) during a line-up task. In keeping with this argument, we also hypothesized that presenting

[1] This was not the first procedure that presented line-up members in pairs. Wells and Pozzulo (2006) examined adult eyewitness' memory of multiple perpetrators using six, two-person line-ups, presented serially. In this procedure, participants were shown pairs of line-up members and were told that each pair may or may not contain a picture of either the assailant or accomplice, but not both. For each pair, participants were required to make a decision before moving onto the next pair. Although not significant, Wells and Pozzulo found a trend suggesting that two-person serial line-ups produced more correct rejections than a simultaneous or sequential line-up.

[2] Note that Pozzulo and Lindsay's (1998) fast elimination line-up does force children to select one face from a series (i.e., more than two) of distractors.

eyewitnesses with pairs of faces will allow for the assessment of discriminability between previously seen and previously unseen faces using a paired-presentation method that is less cognitively taxing on witnesses. Thus, we anticipated that presenting complex stimuli in smaller pieces would reduce the amount of target irrelevant information presented to the witness—thereby reducing the cognitive load required for each decision task. This, in turn, was hypothesized to translate into more accurate responding.

In the present study, we compared the RFC procedure with the face-off procedure (Price & Fitzgerald, 2016) and the simultaneous procedure. Both the RFC and face-off procedures present line-up members in pairs and, as such, both procedures were expected to produce more accurate responding than the simultaneous procedure. However, we also expected to observe differences between the face-off and RFC procedure. Although the face-off and RFC procedures employ similar presentation formats (i.e., paired presentation of line-up members), the RFC is unique in that it introduces a feature of repetition. That is, rather than seeing only 'winning' faces repeatedly (as in the face-off procedure), witnesses will see all faces the same number of times.

The repetition advantage

In the RFC procedure, a witness sees every line-up member paired off with every other member. In an eight-person line-up, this results in a witness seeing each line-up member seven times. Critically, the use of a two-alternative forced-choice model, like the RFC procedure, provides quantitative information about a witness' recognition memory that would normally not be available (Jang et al., 2009). That is, the repetitive nature of the RFC procedure allowed us to collect information about response times at each exposure, selection patterns that rank the target's face relative to faces in the line-up, and the witnesses' final decisions. Thus, the repetition provides an opportunity to collect supplementary information about a witness' memory, without solely relying on his or her final line-up decision (and a possible accompanying confidence assessment). This information can be used to better understand a witness' decision and help investigators attribute weight to a decision.

Present experiments

Two experiments and a supplementary model development analyses were conducted. In experiment 1, we compared children's (aged 6 to 11 years) performance on the RFC procedure with two other line-up procedures: the well-established simultaneous procedure with wildcard (all line-up members shown at once) that has become a standard procedure in the child eyewitness literature against which new procedures are compared (e.g., Zajac & Karageorge, 2009; Price & Fitzgerald, 2016) and the recent face-off procedure (Price & Fitzgerald, 2016). Given that children are frequently tested on simultaneous procedures (cf. sequential; Lindsay & Wells, 1985), we thought an important first step was to compare the new RFC procedure against the simultaneous (with wildcard) procedure. This comparison was especially important because all compared procedures included some level of relative judgment in decisions (i.e., comparing more than one face at a time). In experiment 2, we compared adult witness' performance on the RFC to the simultaneous (with wildcard) procedure. Finally, in the supplementary analyses, we developed a model to better understand individual witnesses' selection bias for the suspect during the experiments 1 and 2 for those who completed the RFC procedure.

EXPERIMENT 1

Developmental differences in eyewitness line-up identification

An important focus of this research was development and the unique circumstances that differently impact child and adult eyewitnesses. It is frequently reported that child eyewitnesses are more likely to choose from a line-up when they should reject the line-up (Fitzgerald & Price, 2015). The cause of children's high rate of choosing is not fully understood, but there has been a great deal of research attempting to understand why this robust age difference exists and to develop strategies to reduce or eliminate children's problematic choosing (e.g., Dunlevy & Cherryman, 2013; Humphries, Holliday, & Flowe, 2012; Parker & Ryan, 1993; Price & Fitzgerald, 2016; Pozzulo & Lindsay, 1998).

Several developmental line-up researchers have introduced procedural advancements in an effort to reduce this problematic choosing. For instance, some researchers have found that adding a visual image to the line-up to represent the absence of the perpetrator provides children with a nonverbal method to reject a line-up and reduces misidentifications (e.g., Dunlevy & Cherryman, 2013; Havard & Memon, 2003; Zajac & Karageorge, 2009). For example, Zajac and Karageorge (2009) were able to reduce children's inaccurate identifications by 25% in TA line-ups (at no cost to TP performance) by including a silhouette shape with a question mark in the line-up (the 'wildcard'). The success of the wildcard demonstrates that, with proper supports, children's performance on line-up tasks can be improved. To determine if the current investigation demonstrates improvement beyond that of salient rejection options, Zajac and Karageorge's (2009) wildcard was included in all line-ups.

There is also some evidence that changing the visual presentation of images in a photo line-up can reduce problematic choosing. The *elimination* procedure was developed by Pozzulo and Lindsay (1999) and breaks the identification task down into two decisions. First, children are asked to choose (from an array), who looks most like the person they saw. Next, all but the selected face is removed, and children are asked to decide if it is, in fact, the person they saw. Pozzulo and colleagues have found evidence that the elimination procedure increases the correct rejection rate with no cost to correct identification rates (Pozzulo & Lindsay, 1999; 3- to 6-year-olds; Pozzulo, Dempsey, & Crescini, 2009; 8- to 13-year-olds; Pozzulo & Balfour, 2006; see also Humphries et al., 2012). Similarly, Price and Fitzgerald (2016) proposed the *face-off* procedure that shows line-up members in pairs in order to reduce the stimulus set size and make decisions more manageable for children.

Relative to simultaneous and elimination line-ups, Price and Fitzgerald (2016) found that the face-off procedure reduced children's problematic choosing when the target was not included in the line-up, and this benefit did not come at the cost of correct identifications. Despite these advancements, children—especially young children—still demonstrate problematic choosing levels (Fitzgerald & Price, 2015).

To explore for age effects in the present research, we examined age categorically, rather than continuously. A recent meta-analysis reported that children start performing more like adults on line-up tasks around the age of 9 years (Fitzgerald & Price, 2015). As children age past 9 years, we can expect to observe a more gradual increase with an eventual plateauing of identification accuracy. Moreover, metacognition and executive functioning is thought to play a role in children's identification performance (Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001; Roebers & Howie, 2003). Previous research has found that around the age of 9 to 10 years, children begin to perform more like adults on metacognitive association tasks (e.g., Schneider, 1986). For this reason, we further expected to observe differences between younger and older children's performance. We anticipated that the RFC procedure would improve older children's performance, relative to the standard, or simultaneous, line-up procedure. For younger children, the direction of our hypotheses was less clear. The RFC procedure may help younger children overcome some of more pronounced developmental limitations that often interfere with a line-up decision (e.g., compressing all line-up information into one, single decision). However, proper use of a decision-intensive procedure, like RFC, may require more developed metacognitive and executive functioning levels that younger children are less likely to possess. Research has demonstrated that children, especially younger children, lack the metacognitive capacity to effectively regulate their recognition memory reporting (Bryce & Whitebread, 2012; Keast et al., 2007; Roebers, 2002; Roebers & Howie, 2003). Moreover, younger children's underdeveloped executive control may make them especially vulnerable to decision fatigue where high numbers of decisions have been found to produce deteriorating decisions or avoidance of decisions (e.g., Vohs et al., 2014).

Method

Participants and design

Children aged 6 to 11 years ($M_{age} = 9.04$, $SD = 1.52$), $n = 451$, were recruited from a summer camp. Participants witnessed a live event in which two people interrupted their activities to search for a lost item. The next day, children were randomly assigned to participate in two identification tasks, one for each of the two targets. This study was a 3 (line-up procedure: simultaneous, face-off, and RFC) × 2 (target: present and absent) × 2 (age: 6- to 8-year-olds and 9- to 11-year-olds) design. The age distributions for each condition were highly similar: simultaneous ($M = 9.00$, $SD = 1.55$, range = 6–11), face-off ($M = 9.05$, $SD = 1.52$, range = 6–11), and RFC ($M = 9.10$, $SD = 1.48$, range = 6–11).

Line-ups

All line-ups contained eight members, including one suspect and seven fillers. In addition, line-ups all contained a visual rejection option (the wildcard; Zajac & Karageorge, 2009). Fillers and innocent suspects (16 photographs and eight for each target) were selected from the Glasgow Unfamiliar Face Database (Burton, White, & McNeill, 2010) and were consistently used across all line-up types. To select the pictures, 200 photographs that matched the two targets' general description as determined by the researchers (gender and hair color) were preselected (100 women and 100 men). Next, independent adult judges ($n = 24$[3]) provided pairwise similarity ratings between photographs of the target and 100 potential fillers (preselected on gender and hair color) on a 10-point Likert-type scale (1 = not at all similar and 10 = highly similar). Mean ratings were used to select fillers to ensure fillers were neither too similar nor too distinct from the target (Fitzgerald, Price, Oriet, & Charman, 2013). Overall, mean ratings were very low that produced little variation in similarity ratings. The mean similarity rating for the female fillers was slightly lower ($M = 3.16$) than for the male target ($M = 3.60$). All line-up images (180 H × 288 W pixels) were displayed on an 11-in. touch screen tablet and shown using EPRIME 2.0 software that recorded participants' responses. Line-up bias measures indicated that the line-ups were appropriate (see details in section 2.1 of the Supporting Information).

Line-up presentation was dependent upon procedure condition, described in the following Procedure section. Note that we did not include the sequential procedure as a comparison procedure as children have performed poorly on this task (Lindsay, Pozzulo, Craig, Lee, & Corber, 1997; Pozzulo & Lindsay, 1998). Moreover, we prioritized using procedures that involved an element of direct relative judgment (i.e., seeing multiple faces at once) in this initial exploration to ensure some comparability across procedures. Target presence was manipulated, with half of line-ups TP and half TA. TP line-ups contained the target from the live event as well as seven fillers. In TA line-ups, targets were replaced with the most similar-looking innocent suspect (similarity rating: $M = 3.32$ for female; $M = 5.86$ for male). The order/location of line-up images, including the suspect, was randomized across participants. The position of the wildcard in each line-up was held constant. The computer recorded children's response latency (time to selection or touching the chosen face) for each line-up decision made.

Procedure

On the first day, a male and a female research assistant visited children in small groups (about 10–15 children per group) and introduced themselves as scientists who had been working in the room before the children arrived. The targets informed the group that they lost something of value and needed the group's help to find it. As a group, they searched for the object. During the search, the targets moved around the room and interacted with many children to ensure they

[3] The $n = 9$ independent adult raters rated similarity for the female target and an additional 15 independent adult raters rated similarity for both female and male targets.

were memorable. Once the male target found the missing object in a silly place (his pocket), they thanked the children and left. The interaction was audio recorded to ensure consistency across groups, and each session lasted about 10 minutes.

The following day, research assistants worked with children individually. Only children with parental consent were invited to participate. Following rapport building (e.g., open-ended questions about camp), children completed the line-up tasks (i.e., randomly assigned into one of three procedures, described in the succeeding texts). A computer program was used to administer the line-up to limit administrator influences (Wells & Bradfield, 1998), although research assistants read the instructions aloud that were visible on the screen to ensure children of all ages received the instructions. Children were told that they would see some pictures and were asked if the visiting scientists who had interrupted their class were in one of the pictures. Children were asked to identify both targets and, as such, were shown two line-ups (one for each target). The order of the line-ups was counterbalanced (half saw the male first, and half saw the female first), but children always completed the same procedure for both targets (e.g., both simultaneous). Participants were told that the people from the videos may or may not be present in the line-up (i.e., nonbiased instructions). The research assistant who launched the computer program did not know whether the suspect would be present in the line-up; the computer randomly determined this.[4] In addition, administrators sat across from the children, holding the tablet computer in such a way that prevented them from seeing the participant's selection. In the following, the description of three line-up procedures and associated instructions are provided.

Simultaneous. Simultaneous line-ups were displayed in a randomized, 3 × 3 array with the wildcard in the center of the array (i.e., position 5 of 9). Participants were asked to either select the line-up member they believed was the target (i.e., Dr. Matthew/Jessica) or indicate that the target was not in the line-up (i.e., select the wildcard) with the following instructions:

> The computer is going to show you some pictures. I need you to help me figure out if Dr. Matthew/Jessica's picture is one of the pictures. There may be a picture of Dr. Matthew/Jessica here or there may not be a picture of Dr. Matthew/Jessica. Think back to what Dr. Matthew/Jessica looked like. I want you to compare your memory of Dr. Matthew/Jessica to each of these pictures. If you see Dr. Matthew/Jessica's picture, touch it. If you don't see his/her picture, touch the shadow picture in the middle.

[4] Target presence was randomly assigned across the different line-ups within each participant. Participants were presented with one of the following orders of line-up conditions for the two line-ups they were shown: (1) TA for first line-up, and TA for second; (2) TA and TP; (3) TP and TA; and (4) TP and TP. This was performed to minimize potential impact of target presence in the first line-up on subsequent decisions. Because this was randomly assigned by computer, we checked to make sure there was no bias or difference in responding (suspect, filler, or rejection responses) between these four groups. No differences emerged in TA or TP conditions ($ps > .05$).

Face-off. Face-off line-ups were presented in pairs. There were four *rounds* presented and, for each round, participants were presented with several pairs. For each pair, participants were asked to identify the person who looked MOST like the target. For round 1, two random faces (pair 1) were compared (i.e., a face-off match). The winner (chosen line-up member) for the first match was stored for later use, and the loser (non-chosen line-up member) was removed from the line-up. This was repeated for the next three pairs (i.e., pairs 2, 3, and 4). During round 2, participants were presented with the winners of pairs 1 and 2, and the winner of this match was stored for the next round while the loser was removed. Next, participants were presented with winners of pairs 3 and 4. During round 3, participants were presented with the two winners from round 2. Round 4 involved showing participants the overall winner and the wildcard and asking them to make a final decision as to whether the winner was the target. Participants were presented with the following instructions:

> [Round 1–3] The computer is going to show you some pictures. I need you to help me figure out if Dr. Matthew/Jessica's picture is shown. There may be a picture of Dr. Matthew/Jessica shown or there may not be a picture of Dr. Matthew/Jessica. The computer is going to show you two pictures at a time, in pairs. Each time you see a pair, I want you to touch the person who looks MOST like Dr. Matthew/Jessica. Even if you don't think either picture looks like Dr. Matthew/Jessica, I need you to touch the one that looks MOST like him/her. Which of these looks MOST like Dr. Matthew/Jessica?

> [Round 4] The computer is going to show you the picture that you thought looked MOST like Dr. Matthew/Jessica, but that doesn't necessarily mean it is Dr. Matthew/Jessica. Remember, his/her picture might not have even been shown at all, so this might be a picture of Dr. Matthew/Jessica or it might be a picture of someone else. If you think it is Dr. Matthew/Jessica's picture, touch the picture. If you don't think it is Dr. Matthew/Jessica's picture, touch the shadow picture.

Repeated forced-choice. Participants in the RFC line-up condition were also presented with line-up members in pairs. This line-up contains three rounds presented to participants and, for each round, participants were presented pairs of faces. For each pair, participants were asked to identify the one who looked MOST like the target (i.e., a most similar judgment). For round 1, participants were asked to decide 28 pairs—that is, they saw all possible combinations of face pairs. During round 2, participants were presented with the two faces they chose MOST often. Again, they were asked to choose the one who looked MOST like the target. For round 2 to be successful, the computer program was designed to deal with situations of 'tied' line-up members by selecting the face chosen during the direct comparison of the 'tied' faces. See sections 1.2 and 1.2.1 in the Supporting Information for detailed information on 'tied' line-up members.

During the final round (round 3), participants were presented with the face they selected from the previous round along with the wildcard. They were asked to make a final decision about whether or not the face was the target. Pilot testing using children of a similar age range was conducted to ensure the functionality of the program as well as to glean feedback from children on task length and comprehensibility (see section 1.1 in the Supporting Information for a full description). Rounds 1 and 3 of the RFC procedure included the same instructions as 'rounds 1–3' and 'round 4' (respectively) of the face-off procedure. In addition, round 2 included the following instructions:

> [Round 2] Now the computer is going to show you the two pictures that you picked most often. I need you to pick the one that you think looks MOST like Dr. Matthew/Jessica.

Once each line-up was completed, children were asked to provide a confidence rating (i.e., *how sure are you in your choice?*) on a visual, numerical (ranging from 1 through 11) on the tablet computers. The following instructions were used, 'Now I would like you to tell the computer how sure you are in your decision by selecting a number on this screen. If you are very sure, you should choose a higher number. If you are not very sure, you should choose a lower number. If you are a little bit sure but not too sure, you should choose a number somewhere in the middle' Confidence ratings were provided prior to any subsequent discussion or interaction with the research assistant. After completing the first line-up, children immediately advanced to the second line-up. Once completed, children were thanked and given a prize.

Results

We conducted several analyses to compare the RFC procedure to the simultaneous and face-off procedures. First, we explored the impact of the procedures on the accuracy of line-up identification decision, followed by a comparison of line-up diagnosticity for these procedures. Next, we explored the predictive and discrimination value of the supplementary memorial information provided by the response patterns of those who completed the RFC procedure.

Identification responses

Children's identification responses were categorized into one of three responses: suspect identifications, filler selections, or line-up rejections. Suspect identifications represent a correct identification in TP conditions and a false identification of the innocent suspect in TA conditions. If the wildcard was selected (i.e., no line-up member was chosen), the response was classified as a rejection. A rejection was correct in a TA condition, but an error in the TP condition. Filler selections were always errors. See Table 1 for the identification response rates for the simultaneous, face-off, and RFC procedures.

Line-up identification choice

To understand the influence of line-up procedure and eyewitness age on identification responses for each of the two targets, 3 (procedure: simultaneous, face-off, and RFC) × 2 (actor: male and female) × 2 (age: 6–8.99 and 9–11.99) × 2 (target presence: present and absent) × 3 (line-up response: suspect, filler, and reject) hierarchical log-linear analyses (HILOG) were conducted with line-up response as the dependent variable. Odds ratios (ORs) and associated 95% confidence intervals (CIs) were computed as an effect size to indicate significant differences between procedures. CIs that do not overlap with 1.00 represent a significant difference ($\alpha = .05$). The HILOG revealed that the highest order interaction (i.e., a five-way interaction between all variables) was nonsignificant, $\chi^2(4) = 3.55$, $p = .47$, indicating that the model containing all five variables did not provide adequate fit with the data, and therefore, the association between all five variables together was not explored further. The highest order effect was a four-way interaction, $\chi^2(16) = 28.66$, $p = .03$. Partial association revealed a significant relationship between age, target presence, procedure, and line-up response, $\chi^2(4) = 11.45$, $p = .02$, as well as a significant interaction between age, procedure, and line-up response, $\chi^2(2) = 9.61$, $p = .008$. To further explore these associations, we examined the relationship between age, procedure, and line-up response separately for TP and TA conditions.

Table 1. Identification responses for younger and older children in experiment 1

Age (years)	Target	Procedure	Identification decision: Suspect	Filler	Reject	*n*
6–8	Present	Simultaneous	0.43	0.18	0.38	60
		Face-off	0.45	0.12	0.43	49
		RFC	0.32	0.33	0.35	46
	Absent	Simultaneous	0.14	0.28	0.58	50
		Face-off	0.04	0.40	0.56	57
		RFC	0.20	0.36	0.44	50
9–11	Present	Simultaneous	0.52	0.17	0.31	94
		Face-off	0.54	0.20	0.26	81
		RFC	0.58	0.19	0.23	108
	Absent	Simultaneous	0.03	0.28	0.69	96
		Face-off	0.12	0.32	0.56	81
		RFC	0.06	0.31	0.63	83
Total	Present	Simultaneous	0.49	0.18	0.34	154
		Face-off	0.51	0.17	0.32	130
		RFC	0.51	0.23	0.27	154
	Absent	Simultaneous	0.07	0.28	0.65	146
		Face-off	0.09	0.36	0.56	138
		RFC	0.11	0.33	0.56	133

Note: Given that each witness made two identifications, *n* represents the number of identifications made in each condition. Nine children withdrew mid-task and completed only one lineup (four did not complete the male target lineup, and five did not complete the female target lineup). RFC, repeated forced-choice.

Target present. In TP conditions, there were no significant interactions between line-up response and procedure for younger children, $\chi^2(4) = 6.48$, $p = .17$, nor for older children, $\chi^2(4) = 1.68$, $p = .79$, indicating similar responding across all three line-up procedures.

Target absent. In TA conditions, there was a trend towards a significant interaction between line-up response and procedure for younger children, $\chi^2(4) = 8.49$, $p = .08$, such that younger children picked the innocent suspect

nearly seven times more often in the RFC procedure than in the face-off procedure, $z = 2.55$, $p = .01$, OR = 6.53 [95% CI = 1.34, 31.66]. There was no statistical difference in responding between the simultaneous and RFC conditions. For older children, the relationship between line-up response and procedure was not significant, $\chi^2(4) = 7.02$, $p = .14$, indicating similar responding across all three line-up procedures. For discussion on performance between younger and older children, see section 2.2 of the Supporting Information.

The HILOG also revealed a two-way interaction between actor and response, $\chi^2(2) = 14.67$, $p = .001$; however, there was no significant three-way interaction between actor, response, and procedure, $\chi^2(4) = 0.61$, $p = .96$. This indicates that the absolute performance was different across the two actors (targets) that can be expected given the high level of variability between different faces. However, these differences did not vary by line-up procedure. As such, the two targets were collapsed for all subsequent analyses. A detailed description of differences between actors and follow-up analyses is available in section 2.3 of the Supporting Information.

Line-up comparison

Diagnosticity ratios. Diagnosticity ratios and inferential CIs (ICIs) (Palmer, Brewer, & Weber, 2010; Tryon, 2001) are presented in Table 2 and allow for comparison across the different line-up procedures (Wells & Lindsay, 1980). A diagnosticity ratio provides evidence regarding the likelihood that the identified suspect is the guilty suspect, assuming a base rate of 50%. A ratio of 1.0 indicates that the two events (i.e., identifying a guilty suspect versus identifying an innocent suspect) are equally likely. Departure from 1.0 indicates differences in the probability of these two events (e.g., 2.0 indicates that children were twice as likely to identify the guilty suspect than the innocent suspect). In Table 2, procedures with CIs that do not overlap with one can be considered diagnostic. Diagnosticity was calculated using suspect identification rates (diagnostic of suspect guilt), as well as filler selection and rejection outcomes (diagnostic of suspect innocence; Wells & Olson, 2002).

To compare suspect diagnosticity across the three procedures, we calculated a z-value ratio using the arcsine transformation method (Wells & Olson, 2002) and report relative risk ratios (RRRs) as an indicator of effect size (Bland & Altman, 1995). For RRR, CIs not overlapping with 1.00 indicate statistical significance. For younger children, the RFC procedure produced the lowest diagnostic value—however, this was not significantly different from either the simultaneous, $z = -1.36$, $p = .91$, $RRR = 0.53$, [0.19, 1.46], or the face-off procedure, $z = -2.85$, $p = 1.00$, $RRR = 0.13$ [0.03, 0.61]. Similarly, the simultaneous and face-off procedures did not differ, $z = -1.54$, $p = .94$, $RRR = 0.24$ [0.05, 1.18]. For older children, the RFC procedure was not significantly more diagnostic than the simultaneous procedure, $z = -0.08$, $p = .53$, $RRR = 0.58$ [0.14, 2.41], but approached significance when compared with the face-off procedure, $z = 1.41$, $p = .08$, $RRR = 2.20$ [0.76, 6.35], while the simultaneous was nearly more diagnostic than the face-off procedure, $z = 1.48$, $p = .07$, $RRR = 3.79$ [1.04, 13.72]. Additional diagnostic information (i.e., information gain) is available in section 2.4 of the Supporting Information.

Supplementary memorial information

So far, we have reported the traditional information collected from line-up procedures: witnesses' line-up selection and what this suggests about a suspect's likelihood of guilt. From these metrics, the RFC procedure performs similarly to the simultaneous line-up procedure with older children, but worse with younger children. To fully explore the utility of the RFC procedure, it is important to consider the other, supplementary information available to contextualize a witness' line-up choice from traditional procedures, including response latency and post-identification confidence ratings. Details of confidence-accuracy relations and response latency analyses are available in the Supporting Information (section 2.5 and 2.6). In addition to confidence and response latency information, the RFC procedure also provides supplementary information that can be used to further index recognition memory. During round 1 of the RFC procedure, witnesses are asked to compare all line-up members to each other and, in doing so, they effectively 'rank' the faces from most to least similar to their memory of the target. We explored whether this pattern of responding could provide a useful index of the degree of match between each picture

Table 2. Diagnosticity ratios and 95% inferential confidence intervals for experiment 1

		Diagnostic of guilt			Diagnostic of innocence					
		Suspect			Filler			Rejection		
Age (years)	Procedure	DR	UL	LL	DR	UL	LL	DR	UL	LL
6–8.99	Simultaneous	3.10	5.52	1.74	1.53	2.63	0.89	1.51	2.03	1.13
	Face-off	12.80	28.23	5.80	3.30	6.31	1.72	1.31	1.74	0.99
	RFC	1.63	2.68	0.99	1.10	1.65	0.74	1.27	1.85	0.87
9–11.99	Simultaneous	16.68	36.50	7.62	1.65	2.60	1.05	2.23	2.84	1.75
	Face-off	4.40	7.06	2.74	1.63	2.42	1.09	2.14	2.91	1.58
	RFC	9.68	19.06	4.92	1.69	2.51	1.14	2.71	3.58	2.05

Note: Inferential confidence intervals were calculated (Tryon, 2001) using a bootstrapping procedure (Palmer et al., 2010) in which log scales (in) were used to more closely approximate a normal distribution. The confidence intervals in Table 2 have been converted from log back to their original unit. For suspect identifications, diagnosticity was computed as the ratio of target-present/target-absent responses. For filler selections and rejections, diagnosticity was computed as the ratio of target-absent/target-present responses. DR, diagnosticity ratio; LL, lower limit; UL, upper limit of 95% inferential confidence interval; RFC, repeated forced-choice.

and the witness' memory of the target. For example, if a line-up member is selected in only one out of seven pairings, it suggests a low degree of match, whereas five picks of seven pairings suggests a higher degree of match. To simplify response patterns, we calculated indices (i.e., standard score or *Z*-scores) that summarized information from the entire distribution of responding for each witness with respect to their selection of the suspect. Detailed instructions of how *Z*-scores were calculated are provided in the Appendix. We then grouped *Z*-scores and compared how often each response pattern aligned with each line-up outcome (Table 3).

For both younger and older children, *Z*-scores were significantly higher when the witness selected the guilty suspect (younger: $M = 1.35$, $SD = 0.35$; older: $M = 1.53$, $SD = 0.22$) than when they selected the innocent suspect (younger: $M = 0.85$, $SD = 1.00$; older: $M = 1.08$, $SD = 0.54$) (younger: $t(23) = -1.84$, $p = .004$; older: $t(66) = -3.85$, $p = .01$). When children correctly chose from the line-up, younger children had significantly lower *Z*-scores than older children, $t(76) = -2.58$, $p = .01$, indicating that young children were less able to discriminate guilty from innocent suspects in the line-up. When children incorrectly rejected the line-up, younger children had significantly lower *Z*-scores ($M = -.12$, $SD = 0.79$) than older children ($M = 0.46$, $SD = 0.80$), $t(39) = -2.28$, $p = .03$. No other decisions produced different *Z*-scores. Similar results of a positive relationship between increasing *Z*-scores and accuracy are depicted through a profile analyses of discrepancy associated with *Z*-scores (see section 2.7.1 of the Supporting Information). Thus, higher *Z*-scores were related to accuracy.

Discrimination. To guide interpretation of individual witness responses, we used information about witness' response patterns during round 1 to calculate the adjusted normalized discrimination index (ANDI) that indicates how well participants' selection patterns discriminated guilty suspects from fillers and innocent suspects (for the formulae, see Yaniv, Yates, & Smith, 1991). ANDI was calculated using raw pick rates (e.g., face was picked four times) to calculate how frequently the previously seen face (i.e., guilty suspect) won (out of seven views) relative to the previously unseen faces. ANDI is a measure of variance in accuracy accounted for by patterns of responding that range from 0 (no discrimination) to 1 (perfect discrimination). For example, an ANDI score of .30 indicates that the response pattern can explain 30% of the variability in outcomes. A bootstrapping procedure was used to compute 95% ICIs[5] around the ANDI scores. Younger children's pattern of responding during the first 28 decisions of the RFC procedure do not discriminate the target from unseen faces (ANDI = 0.01, 95% CI [−0.01, 0.03]). Older children were significantly better able to discriminate the target using the RFC procedure (ANDI = 0.23, 95% CI [0.18, 0.28]).

Table 3. Frequency of *Z*-scores for each decision type for each age group in experiment 1

	TA			TP		
Z-statistic	Suspect	Filler	Reject	Suspect	Filler	Reject
Younger children	$n = 6$	$n = 21$	$n = 23$	$n = 14$	$n = 16$	$n = 16$
2.00–2.47	0.00	0.00	0.00	0.00	0.00	0.00
1.75–1.99	0.00	0.00	0.04	0.07	0.00	0.00
1.50–1.74	0.33	0.05	0.09	0.43	0.00	0.06
1.00–1.49	0.50	0.14	0.09	0.43	0.19	0.06
0.00–0.99	0.17	0.05	0.13	0.07	0.44	0.19
−2.47 to (−0.01)	0.00	0.76	0.65	0.00	0.38	0.69
Older children	$n = 4$	$n = 26$	$n = 53$	$n = 63$	$n = 20$	$n = 25$
2.00–2.47	0.00	0.00	0.00	0.05	0.00	0.00
1.75–1.99	0.00	0.00	0.02	0.08	0.00	0.00
1.50–1.74	0.50	0.04	0.08	0.46	0.10	0.08
1.00–1.49	0.50	0.04	0.13	0.40	0.25	0.24
0.00–0.99	0.00	0.39	0.40	0.02	0.45	0.44
−2.47 to (−0.01)	0.00	0.54	0.38	0.00	0.20	0.24

Note: Columns may not sum to one due to rounding. *Z*-score bins were selected for ease of interpretation. TA, target absent; TP, target present.

Predictive utility of Z-scores. To assess whether knowledge of a witness' pattern of responding throughout the RFC procedure allows for post-diction of accuracy, we conducted logistic regressions to assess the relationship between *Z*-scores[6] and identification accuracy. Given the demonstrated role of post-identification confidence ratings in the literature, it was important to learn whether *Z*-scores predict accuracy beyond the confidence ratings. Table 4 outlines the model results. When examining the overall accuracy of responding, inclusion of *Z*-scores (step 2) was predictive of accuracy over and above the inclusion of confidence ratings for older children; however, no model was produced for younger children. Adding *Z*-scores into the model containing confidence ratings increased the explanatory power of the model by about 7%, accounting for about 21% of the variance. This 21% accuracy variance in the RFC procedure is much higher than the 5% of variance accounted for by confidence ratings alone in the simultaneous line-up ($B = .18$, $\text{Exp}(B) = 1.19$, $p = .01$; Nagelkerke pseudo $R^2 = .05$, $\chi^2(1) = 6.95$, $p = .008$; see section 2.5 of the Supporting Information). Next, we conducted separate logistic regressions for positive (i.e., those who chose from the line-up) and negative (those who rejected the line-up) decisions for each age group. Neither *Z*-scores nor confidence explained variance in negative response accuracy. For positive response accuracy, only *Z*-scores were included in the model and accounted for 55% of the variance in accuracy for younger children. For older children, *Z*-scores accounted for 77% of the variance in accuracy and when combined with confidence ratings (step 2), account for 80% of the variance for older children.

Next, using a predictive equation ($\text{Log}(p/1 - p) = b0 + b1 \times 1 + b2 \times 2$), we calculated and plotted the probability estimates of the model at each level of confidence for both overall accuracy and positive response accuracy for older children. As seen in Figure 1, witnesses who provided a confidence rating of 7 or higher had a higher than chance (.50) probability of an accurate decision. Between

[5] This procedure (Palmer et al., 2010; Tryon, 2001) used the observed data as a sampling distribution and conducted 250 replications to estimate variance of ANDI. This estimated variance provided the distribution needed to calculate CIs.

[6] Because of high correlations between number of suspect picks and *Z*-score ($r > .97$, $p < .001$ for both age groups), the model produces comparable results when number of suspect decisions are included in the model in place of *Z*-scores.

Table 4. Logistic regression model information: predictive utility of confidence ratings and *Z*-scores

	Step 1				Step 2			
	B(*SE*)	Exp(*B*)	R^2	χ^2	*B*(*SE*)	Exp(*B*)	R^2	χ^2
Positive								
Younger								
Z-score	3.41(1.34)	30.29	.55	24.66*				
Older								
Z-score	5.41(1.18)	224.54	.77	96.75 *	5.36(1.24)	213.22		
Confidence					0.49(.20)	1.64	.80	104.16*
Total (older)								
Confidence	.34*(.08)	1.41	.14	20.76*	0.31*(.08)	1.36		
Z-score					0.54*(.16)	1.71	.21	32.07*

Note: R^2 refers to Nagelkerke R^2. Degrees of freedom associated with chi-squared tests are 1 for step 1 and 2 for step 2. No model for total accuracy was produced for younger children. *Indicates statistical significance at a level of $p < .001$.

confidence ratings of 7–9, including the witness' *Z*-score with the confidence rating increased the probability of accurate responding. Beyond a confidence rating of 9, *Z*-scores did not notably add to the predictive utility of confidence ratings. In addition, *Z*-scores greater than 1.00 add to the predicted probability of accuracy for positive responses—particularly during the central confidence ratings (e.g., 4 through 7) where there may be some apprehension about interpreting the predictive utility of the ratings.

To gain a better understanding of how confidence and *Z*-scores interact to predict accuracy, see Figure 2. Regardless of the level of confidence provided, a *Z*-score of .50–.99 is not predictive of accuracy. *Z*-scores lower than .50, when paired with high confidence ratings (e.g., 10) are somewhat indicative of accuracy. In particular, *Z*-scores higher than 1.00 helped to predict accuracy when corresponding levels of confidence might otherwise be considered ambiguous (i.e., post-identification confidence ratings of 4 through 7).

Discussion

Experiment 1 examined the impact of the RFC line-up procedure on younger (6- to 8-year-olds) and older (9- to 11-year-olds) children's line-up performance, relative to a simultaneous and face-off procedure.

Identification accuracy

When considering the impact of the RFC procedure on line-up behavior, age mattered. The RFC procedure was less effective for younger than older children, as indicated by higher innocent suspect selections. The face-off procedure appears to have worked best for younger children, with that procedure showing higher diagnosticity than the other procedures. In addition, the face-off procedure reduced young children's overall levels of choosing when compared with the RFC, although this reduction did not appear to align with better discrimination of guilty from innocent suspects. It is particularly interesting that the face-off procedure produced

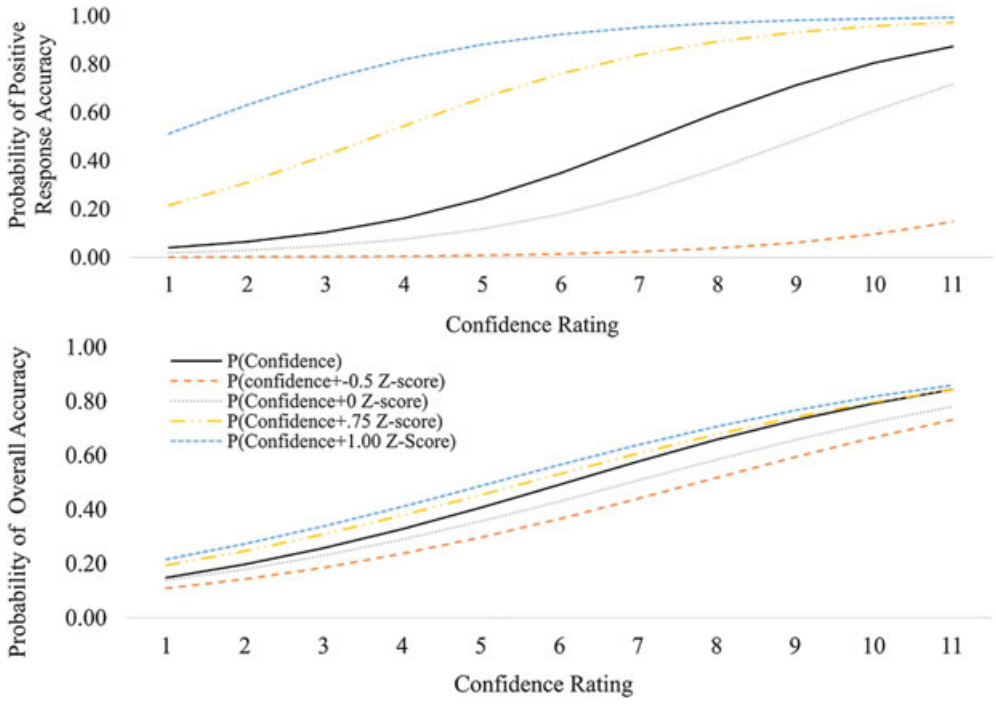

Figure 1. Plotted probabilities of overall and positive response accuracy at each level of confidence for older children. [Colour figure can be viewed at wileyonlinelibrary.com]

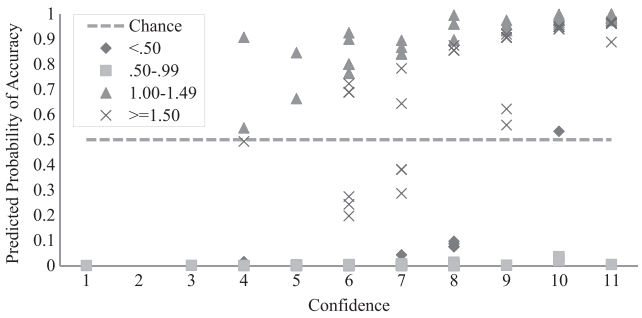

Figure 2. Predictive probability of *Z*-scores and confidence ratings for older children

more accurate responding than the simultaneous with wildcard procedure because the latter procedure has previously worked well with children in this younger age group (e.g., Karageorge & Zajac, 2011). The present data are the first replication of the success of the paired presentation of the face-off procedure reported by Price and Fitzgerald (2016). The high diagnosticity ratios associated with the face-off procedure suggest that, for younger children, the paired-presentation method may be particularly advantageous.

However, not all paired-presentation methods work for young children. Given the presentation similarities between the face-off and the RFC procedures (i.e., dual-presentation/forced-choice decisions and instructions), lower performance during the RFC suggests that younger children struggled with either (a) the length of time to complete the RFC procedure, (b) the large number of decisions, or (c) the repeated exposure to faces. These differences are important to consider for future research. Given that an RFC procedure was completed quite quickly (M = 3.87 minutes per RFC line-up), young children's difficulties with the RFC likely stem from one or both of the latter two possibilities. As the number of decisions increased (i.e., as young children moved through the 28 decisions), the speed of their decisions increased much more than older children or adults (Figures S2 and S4), indicating that the high number of decisions may have resulted in decision fatigue for younger children. In addition, the low ANDI scores for younger children may suggest that repeated exposure to faces decreased children's ability to discriminate previously seen from unseen faces. However, this is an empirical question for future research.

Older children, conversely, performed well with the RFC procedure—at least as well as they performed with the simultaneous (with wildcard) procedure. Both line-up identification decisions and diagnosticity were statistically comparable between the two procedures. Given the strong evidence to support the use of the simultaneous with wildcard procedure with 8- to- 11-year-old children (e.g., Zajac & Karageorge, 2009), the observation that the RFC procedure produced similar responding with a similar age group is encouraging.

Reducing the stimulus set size to two images per decision was hypothesized to reduce the cognitive effort involved in a line-up decision. This, in turn, was hypothesized to benefit children of all ages. However, it appears that the RFC procedure did not allow children to overcome their developmental limitations. Instead, the results suggest that use of the RFC procedure may have required more developed executive control and metacognitive skills that the younger age group (between 6 and 8 years old) have not yet achieved. For instance, notable improvements in inhibitory control have been documented between the ages of 6 and 9 years, wherein children begin to demonstrate improved suppression of automatic or prepotent responses (e.g., Williams, Ponesse, Schachar, Logan, & Tannock, 1999). Similarly, there is evidence that around the age of 9 years, children are better able to monitor the accuracy of retrieved memories and strategically regulate the reporting of memories to improve accuracy than their younger counterparts (e.g., Koriat et al., 2001). Thus, younger children may not have had the executive or metacognitive functioning ability needed to effectively regulate their responding over the large number of decisions.

Prediction of accuracy

Traditional line-up procedures collect one key piece of information from an eyewitness: who, if anyone, is the culprit. If recorded, two other pieces of evidence can be used to contextualize or support the utility of a witness' identification decision: post-identification confidence ratings and response latency. In addition to the post-identification confidence ratings and response latency, the RFC also provided information about witnesses' patterns of responding (i.e., *Z*-scores). Alone, we found evidence that *Z*-scores were indicative of discrimination between guilty and innocent suspects as well as identification accuracy. However, to appropriately evaluate the RFC procedure, it is important to consider witnesses' identification decisions within the context of confidence ratings, as these are frequently collected during line-up procedures. When using *Z*-scores along with post-identification confidence ratings, older children's patterns of choosing are predictive of identification accuracy—above the information provided by confidence ratings. The *Z*-scores were especially helpful in predicting accuracy in situations where confidence was more ambiguous (e.g., confidence ratings of 4 through 7).

Overall, in experiment 1, the RFC procedure emerged as a potentially viable method to extract recognition evidence from 9- to-11-year-old children. Specifically, the results suggest that the RFC procedure performs as well as a standard, simultaneous line-up (with wildcard) procedure in terms of accuracy, and the additional information captured during the procedure adds to the predictive utility of the RFC procedure.

Given that children begin performing more like adults on line-up tasks at around the age of our older group in experiment 1 (age 9 to 11 years; Fitzgerald & Price, 2015) and we observed clear evidence of differential effectiveness across development for the RFC, in experiment 2, we sought to extend the age range of our witnesses into young adulthood.

EXPERIMENT 2

Method

We recruited 153 undergraduate students (M_{age} = 20.86, SD = 5.32; 79% women; 64% White and 10% Black) who received partial course credit as compensation. The same line-up materials from experiment 1 were used, with one

exception—experiment 2 did not include the face-off procedure, a procedure designed for young children that has not previously been used with adults. To allow for a clear comparison with experiment 1, we included the wildcard in both line-up procedures. This study was a 2 (line-up procedure: simultaneous and RFC) × 2 (target: present and absent) between-subjects design.

Procedure

The study was advertised as an investigation of perceptions of personality. Participants individually viewed two target event videos (approximately 2 minutes in length each), each containing one of the same targets as in experiment 1. On the first day, participants were told that they would be watching videos of people and, the following day, would later be asked to answer questions about their perceptions of the people's personalities. In the female target video, the target reads a book, spills a glass of water, and cleans it up. In the male target video, the target is studying and does stretches/exercises. The next day, the experimenter explained the true purpose of the study prior to administering the line-up on a computer program. Participants were asked to complete a second consent form that outlined the true nature of the study. Once consent was given, the experimenter began the line-up procedure.

The experimenter was responsible for launching the computer program and providing instructions but was not near the participant when she or he viewed the line-up or made a decision. The same programs used in experiment 1 were used to administer the line-ups to the adults, with some adjustments to the instructions to better reflect the context (e.g., any references to camp and the visitors were removed; the targets were referred to as the man and the woman in this experiment). After each line-up decision was made, participants were asked to make a confidence judgment using the same scale described in experiment 1. A screening question was used to ensure that participants were unfamiliar with the targets.[7] Experimenters were blind to target presence, but were responsible for ensuring counterbalancing of the order in which the target line-ups were shown (i.e., male or female target first). Participants were told that the people from the videos may or may not be present in the line-ups.

Results

To examine how the RFC procedure compared with the simultaneous procedure with adult eyewitnesses, we first explored the impact of the simultaneous and RFC procedures on the accuracy of line-up identification choice, followed by a comparison of line-up diagnosticity for these two procedures. Next, we explored the value of the supplementary memorial information provided by the response patterns of those who completed the RFC procedure.

Line-up identification choice

To understand the influence of line-up procedure on identification responses, a 2 (procedure: simultaneous and RFC) × 2 (actor: male and female) × 2 (target presence: present and absent) × 3 (line-up response: suspect, filler, and reject) HILOG revealed that the highest order interaction was a two-way effect, $\chi^2(9) = 77.40$, $p < .001$ (Table 5). Partial association analyses indicated a two-way interaction between target presence and line-up response, $\chi^2(2) = 69.97$, $p < .001$, such that more suspect identifications were made in TP line-ups, while more filler and rejections were made in TA line-ups (all ps $< .05$). There was also an interaction between actor and response, $\chi^2(2) = 14.40$, $p = .001$, such that more correct identifications and fewer correct rejections were made in the female than the male line-up. Similar to experiment 1, the lack of a three-way interaction between actor, response, and procedure suggests that the differences in responding across actors were consistent across the two line-up procedures. As such, targets were collapsed for the remaining analyses. For a complete breakdown of responses by actor, see Table S2.

Table 5. Identification responses for experiment 2

Target	Procedure	Identification response: Suspect	Filler	Reject	*n*
Present	RFC	0.71	0.08	0.22	75
	Simultaneous	0.64	0.15	0.21	74
Absent	RFC	0.18	0.36	0.47	73
	Simultaneous	0.23	0.28	0.49	74

Note: Each witness made two identifications. *n* represents the number of identifications made in each condition. Four witnesses did not complete the lineup for the male target as they indicated the target was familiar to them (via screening questions). RFC, repeated forced-choice.

Line-up comparison

Diagnosticity ratios. Diagnosticity ratios were calculated for all three possible decision outcomes for the simultaneous and RFC procedures. As seen in Table 6, the diagnosticity ratios produced by the simultaneous and RFC procedures were not statistically different for all three decisions. Wide inferential ICs indicate that differences were not statistically significant, suggesting that adults had similar responding patterns across the two line-up procedures. The RFC procedure was not significantly more diagnostic of suspect guilty than the simultaneous procedure, $z = 1.20$, $p = .11$, $RRR = 1.42$ [0.72, 2.79]. See section 3.1 of the Supporting Information for additional diagnostic information (i.e., information gain).

Table 6. Diagnosticity ratios and 95% inferential confidence intervals for experiment 2

	Diagnostic of guilt: Suspect			Diagnostic of innocence: Filler			Rejection		
Procedure	DR	LL	UL	DR	LL	UL	DR	LL	UL
Simultaneous	2.76	1.96	3.90	1.91	1.16	3.14	2.25	1.54	3.30
RFC	3.97	2.68	5.88	4.45	2.32	8.54	2.18	1.51	3.16

Note: Inferential confidence intervals were calculated as in experiment 1. DR, diagnosticity ratio; LL, lower limit 95% inferential confidence interval; RFC, repeated forced-choice; UL, upper limit of 95% inferential confidence interval.

[7] Four witnesses did not complete the line-up for the male target as they indicated the target was familiar to them (via screening questions).

Supplementary memorial information

Information pertaining to the confidence-accuracy and response latency-accuracy analyses can be found in the Supporting Information (sections 3.2 and 3.3). Similar to experiment 1, we examined the extent to which *Z*-scores could be used to further index recognition memory. We calculated and grouped *Z*-scores to compare how often each pattern of responding aligned with each line-up outcome. As outlined in Table 7, for all age groups, *Z*-scores that fall between 0 and 1.74 are indicative of suspect identifications (both guilty and innocent suspects), whereas lower and negative *Z*-scores are related to filler identifications and rejections. Unlike what was observed in experiment 1, *Z*-scores were similar for guilty ($M = 1.11$, $SD = 0.47$) and innocent suspect identifications ($M = 1.12$, $SD = 0.64$), $t(63) = 0.03$, $p = .25$.

Discrimination. Adjusted normalized discrimination index scores indicated that adults were able to discriminate the target well using the RFC procedure (ANDI = 0.15, 95% CI [0.08, 0.21]). ICIs allow for the comparison across the two experiments and indicate that the ANDI scores for the younger children were significantly lower than the older children and adults. Although not significantly different, it is worth noting that the RFC procedure accounted for more variance in accuracy (23%) for older children than it did with adults (15%).

Predictive utility of Z-scores. Logistic regressions revealed that confidence ratings following each identification significantly predicted overall accuracy for the simultaneous, $B = .31$, $\text{Exp}(B) = 1.37$; Nagelkerke pseudo $R^2 = .16$, $\chi^2(1) = 18.98$, $p < .001$, and RFC procedure, $B = .30$, $\text{Exp}(B) = 1.35$; Nagelkerke pseudo $R^2 = .13$, $\chi^2(1) = 15.55$, $p < .001$. *Z*-scores or patterns of responding were not predictive over and above the 13% of variance in accuracy explained by post-identification confidence ratings ($B = .31$, $\text{Exp}(B) = 1.36$, $p = .003$; Nagelkerke pseudo $R^2 = .13$, $\chi^2(1) = 13.96$, $p < .001$).

Discussion

The purpose of experiment 2 was to explore how the RFC procedure influenced adults' identification accuracy. Similar to older children in experiment 1, the RFC procedure produced similar choosing and selection behavior as the simultaneous line-up (with wildcard) in adults. Although post-identification confidence ratings predicted overall accuracy, unlike with older children, *Z*-scores (or choosing patterns) did not add to the predictive utility of the procedure in adults. Despite this, ANDI scores suggest that *Z*-scores provided by adults were a useful tool for assessing discriminability between previously seen and unseen faces.

Together with the accuracy rates, diagnosticity ratios, and discrimination scores (i.e., ANDI), these results support the hypothesis that adult witnesses can use the RFC procedure to discriminate between guilty and innocent suspects. Thus, there appeared to be no cost to administering the RFC, relative to the simultaneous line-up (with wildcard). It is important to highlight, however, that the simultaneous procedure in experiment 2 included a visual, salient rejection option (i.e., the wildcard; Zajac & Karageorge, 2009). Although typically used with children, there is evidence to suggest that the wildcard encourages conservative responding from adult witnesses (Bruer, Fitzgerald, Therrien, & Price, 2015). Two implications of this finding should be highlighted. First, considering the similar rejection decisions across the two line-up procedures, it is plausible that the RFC procedure also encourages more conservative decision-making. Second, comparison with a simultaneous-without-wildcard line-up may have produced significant benefits of the RFC procedure. The next natural step in this research is to explore how the RFC compares with more commonly used procedures with adult witnesses, such as the simultaneous procedure (without wildcard) and the sequential procedure.

In sum, there was no disadvantage to asking an adult witness to identify a target using the RFC procedure. This is encouraging because the RFC procedure was designed to capture more information about a witness' memory than other line-up procedures. To further explore the utility of this additional information for interpreting a witness' line-up decision, additional analyses were conducted using data from experiments 1 and 2.

Table 7. Frequency of *Z*-scores for each decision type in experiment 2

	TA			TP		
Z-statistic	Suspect	Filler	Reject	Suspect	Filler	Reject
Adults	$n = 13$	$n = 26$	$n = 34$	$n = 53$	$n = 6$	$n = 17$
2.00–2.47	0.00	0.00	0.00	0.00	0.00	0.00
1.75–1.99	0.00	0.00	0.03	0.04	0.00	0.00
1.50–1.74	0.23	0.12	0.09	0.23	0.00	0.18
1.00–1.49	0.54	0.38	0.29	0.47	0.33	0.29
0.00–0.99	0.23	0.31	0.35	0.15	0.50	0.29
−2.47 to (−0.01)	0.00	0.19	0.24	0.11	0.17	0.24

Note: Columns may not sum to one due to rounding. Unlike the bins used to examine model fit (i.e., grouping *Z*-scores based on third percentile), these bins of *Z*-scores were decided considering ease of interpretation. TA, target absent; TP, target present.

EXPERIMENTS 1 AND 2: MODEL DEVELOPMENT

In experiment 1, and to a lesser extent experiment 2, we outlined the predictive utility of *Z*-scores in estimating witness accuracy in the RFC line-up procedure. That is, calculating an individual witness' *Z*-score can be used alone or in combination with a witness' post-identification confidence rating to estimate the likelihood of memory accuracy. However, the value of a witness' pattern of responding may not be limited to interpreting a single *Z*-score. The repetitive and forced-choice nature of the RFC procedure provides an opportunity to learn more about a witness' recognition memory. The goal of these analyses was to explore the additional information about a witness' target memory provided by choosing patterns during the RFC procedure. To do so, we developed a model that uses an individual witness' pattern of responding as an indication of suspect selection bias that, in turn, may be construed as a proxy for memory strength. In the following, we describe our rationale for the model, how

the model was developed, and how the model aids in interpreting line-up behavior.

Problems with traditional analyses models

Like other alternative line-up procedures (e.g., Sauer et al., 2008), the RFC procedure was developed to provide granular-level data about a witness' decisions. However, unlike other procedures, the round-robin design of the RFC results in challenges that limit the application of previous used statistical analyses (e.g., the profile analyses of Brewer et al., 2012; the classification algorithms of Sauer et al., 2008. In the RFC, each line-up member is compared against every other line-up member. As a witness makes these 28 decisions,[8] there is a degree of mutual influence that each selection will have on the possible remaining scores that can be assigned to each line-up member. For instance, consider a situation in which the first pair of line-up members shown to a witness includes line-up members 4 and 2. When the witness selects the winner (e.g., 2 is most similar to their memory for the target), the winning member has a 'score' of 1, while the losing member has a 'score' of 0. Thus, going forward, the winning member has potential to 'score' the maximum (7 out of 7), while the losing member cannot exceed a score of 6 (out of 7). This interdependency of responding limits the variability that can be observed between line-up members—making it difficult to find meaningful differences using many previously applied approaches. To overcome this, we created a model for use with a round-robin design. Specifically, because the RFC model has a finite number of possible patterns of responding across the eight line-up members, we used these finite patterns to further understand a witness's memory in the form of response probabilities.

A round-robin model using the probability of responses

When a witness chooses from the RFC procedure, his or her pattern of choosing reflects a selection *bias* for the suspect that may be interpreted as proxy for memory strength or discrimination for the suspect. A low selection bias may result if the witness has a weak memory for the suspect (e.g., poor encoding or target is absent from the line-up), and the corresponding response pattern should be more likely to be relatively random across the eight line-up members and reflect little selection bias for any one line-up member. That is, it can be expected that the likelihood of any one line-up member being selected over any other is around 50%. With a stronger suspect selection bias, a witness' pattern of responding should be less random across the eight line-up members. That is, the likelihood that the suspect will be selected over another line-up member should be greater than chance (i.e., 60%, 70%, 80%, or 90%). Importantly, a strong selection bias does not necessarily equate with memory strength for the guilty suspect, but rather could also be a result of mistaken recognition or random commitment effect to the suspect, given that witnesses were instructed to 'select the one that looks MOST like the target'.

To create a tool for applied use, we developed a model that was not based on any single data set, including the present data, but rather based on this conceptual framework of selection bias surrounding the RFC procedure. First, we selected five levels of the model parameter (i.e., *suspect selection bias*) to represent in the model.[9] The first level (50%) reflects a random level of selection bias whereby any line-up member, including the suspect, has a 50% chance of being selected over any other line-up member. For the second level, the model was designed such that the suspect would win over any other face 60% of the time, while each filler continued to have a 50% chance of winning when paired against any other filler. We continued to increase the rate at which the suspect would win for the third level (70%), fourth level (80%), and fifth level (90%). For each of these five selection bias levels, we ran 300 000 simulations to determine the frequency at which all possible[10] patterns of responding occurred. To simplify these model-generated patterns of responding, we calculated *Z*-scores (i.e., indicator of individual witness' discrimination for the suspect) the same way as described in experiments 1 and 2.

Application of model data to experiment data

Model fit

Next, we examined how well the model fit the data from experiments 1 and 2 using a *G*-test for goodness of fit. *G*-statistics for different levels or model parameters (bias between 50% and 90%) were calculated separately for TA and TP conditions, with a good-fitting model indicated by a non-significant difference between the observed data and the expected (model) data. Overall, the model provided adequate fit for experiments 1 and 2 data. For all children,[11] the results indicated that the 48–50% parameter model was a good fit for *Z*-scores for TA conditions (G-statistic >8.28, $p > .04$; Table S7). Thus, when the target is absent, the response patterns are not statistically different from the model's definition of random responding. Likewise, for TP conditions, the results indicate a good fit between experiment 1 data and the 70% parameter model data (G-statistic = 1.40, $p = .71$). For adults, there were similar patterns in TP line-ups, wherein the 70% parameter model provided the best fit for the experiment 2 data (G-statistic = .63, $p = .65$). However, when the target was absent, adults' choosing patterns are best described by the 65% parameter model [cf. children's 50% bias; (G-statistic = 0.29, $p = .96$)].

Taken together, these results suggest that when the target is absent from the line-up, children's behavior reflects random or unbiased choosing. Adults, on the other hand, appeared to more closely (or were better able to) follow the instruction to *select the one that looks MOST like the target* and, in turn, they favored the innocent suspect more than

[8] The formula to determine the number of decisions based on line-up size is $n * (n - 1)$ 2. For example, a line-up containing six members would include 15 decisions (6 * 5 2).

[9] Note that, because of computational intensity of the process, these model parameters were selected by the authors. With more computational power, one could optimize and test the model using 300 000 simulations at more granular increments (e.g., 81, 82, and 83).

[10] All possible patterns when disregarding the rankings/order of fillers win.

[11] Model fit analyses were initially run separately for each age group but produced similar results. For brevity, data were collapsed across age for subsequent analysis of model data. Age-specific model fit data can be found in Table S7.

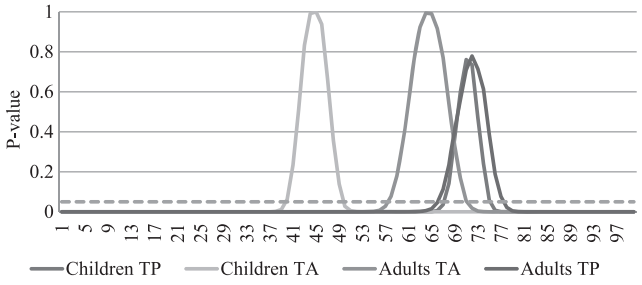

Figure 3. The *p*-values plotted at each level of bias for children (collapsed across age) and adults. Model data were examined using a bootstrapping procedure with 5000 simulations. Note that the data used in the calculations of *G*-statistics are based on results from a sample of 300 000 simulations. In this instance, the number of simulations was reduced (to a level that produced results similar to the model with 300 000 simulations) to allow for more granular examination of each level of bias. The dashed line indicates the level of significance ($p = .05$). TA, target absent; TP, target present

the fillers. See Figure 3 for a visual comparison of which level of selection fit best with each age group. There are several interesting observations to be made in Figure 3. Specifically, the peaks of the TP curves are lower on the *y*-axis (lower *p*-values) than the TA curves. That is, the TP curves do not reach a *p*-value of 1 (i.e., perfect fit). This suggests that, overall, the model provides a better fit for TA data.

Using the model to interpret witness responding

Once we established that the model fit (and which parameter best fit) with experiments 1 and 2 data, we next applied the model data to help interpret individual witness responses from experiments 1 and 2. To do this, we first calculated ORs that indicated the likelihood of strong or weak suspect selection bias associated with a range of *Z*-scores—using only the model data to create a *look-up table* (a partial lookup table is shown in Table 8; the full lookup table is available in Table S8). These ORs in the look-up table indicate the likelihood that the pattern came from a higher suspect selection bias simulation (e.g., 70%) rather than a random (i.e., 50%) simulation.

Next, we were able to compare an individual witness' *Z*-scores to the look-up table to determine the likely level of suspect selection bias demonstrated by that particular witness. Note that this comparison allows only for consideration of suspect (innocent or guilty) bias, not for another line-up member that may have been ultimately chosen from the line-up. For example, take two older child witnesses from experiment 1. These witnesses appear very similar at first glance. Both witnesses correctly identified the guilty suspect, had similar response latency (just over 2 minutes), and provided a confidence rating of 10 (out of 11). The information about *Z*-scores and the probability of selection bias can help to differentiate these similar witnesses. The first witness' pattern of responding translated into a *Z*-score of 1.82. Using the table, we observed that this particular witness' pattern is likely to reflect a selection bias value of 70% or higher (this witness' *Z*-score or selection pattern is 5.47 times more likely in the 70% bias level than in the random bias level). Conversely, the second witness' pattern of responding produced a *Z*-score of 0.94. From the look-up table, we can

Table 8. Model simulation odds ratio: *Z*-statistic look-up table (partial)

			70%		
Z-statistic	OR: 50%	OR: 60%	OR: 70%	OR: 80%	OR: 90%
2.00–2.47	6.93	2.41	—	0.47	0.25
1.75–1.99	5.47	2.16	—	0.52	0.31
1.50–1.74	3.85	1.80	—	0.59	0.52
1.00–1.49	2.54	1.43	—	0.90	1.21
0.00–0.99	1.07	0.91	—	1.55	4.78
−2.47 to (−0.01)	0.26	0.45	—	3.73	47.90
			60%		
Z-statistic	OR: 50%	OR: 60%	OR: 70%	OR: 80%	OR: 90%
2.00–2.47	2.87	—	0.41	0.20	0.10
1.75–1.99	2.54	—	0.46	0.24	0.14
1.50–1.74	2.14	—	0.55	0.33	0.29
1.00–1.49	1.78	—	0.70	0.63	0.84
0.00–0.99	1.18	—	1.10	1.72	5.28
−2.47 to (−0.01)	0.59	—	2.24	8.37	107.51
			50%		
Z-statistic	OR: 50%	OR: 60%	OR: 70%	OR: 80%	OR: 90%
2.00–2.47	—	0.35	0.14	0.07	0.04
1.75–1.99	—	0.39	0.18	0.10	0.06
1.50–1.74	—	0.47	0.26	0.15	0.13
1.00–1.49	—	0.56	0.39	0.35	0.47
0.00–0.99	—	0.85	0.94	1.46	4.49
−2.47 to (−0.01)	—	1.69	3.79	14.14	181.64

Note: OR, odds ratio.

observe that this witness' pattern of responding reflects more random selection bias (this witness' *Z*-score is just as likely to appear in the 50% bias level as the 60% and 70% bias levels, as indicated by OR close to 1). From this, we can surmise that the first witness likely had a stronger selection bias towards the suspect (and perhaps stronger memory for the suspect) than the second witness. This information allows us to understand more about a witness's ability to discriminate (i.e., selection bias) a suspect from fillers—which can be especially useful in situations in which the guilt of the suspect is unclear. Moreover, these ORs can be used to understand a witness' memory for the suspect, even if the final decision was to select a filler or reject the line-up. Take, for instance, a 6-year-old participant who rejected a TP line-up but produced a *Z*-score of 1.69, or an 11-year-old participant who selected a filler from a TP line-up and produced a *Z*-score of 1.50. The table suggests that these patterns are 3.85 times more likely in the 70% bias level than the 50% level. Thus, despite rejecting the guilty suspect or selecting a filler, the selection patterns indicate nonrandom discrimination of the guilty suspect.

The model information presented thus far is designed to only consider information from the first round of the RFC. There are both advantages and disadvantages to this approach. An advantage of relying only on the data provided round 1 as an index of recognition memory is that it removes the reliance on a witness' ability to make a final decision. Not asking a witness to make that final decision may be especially useful for child witnesses who may not have the metacognitive ability to accurately make categorical line-up

decisions (Keast et al., 2007). In addition, using round 1 to build the model does not require knowledge of whether the suspect is guilty or innocent—making the information particularly useful for applied situations when these data are not available. A disadvantage of building a model from only round 1 data is that it maximizes the chance of an innocent suspect identification in situations where the innocent suspect most closely resembles the culprit (biased line-ups). If participants follow the instructions of the RFC procedure, they will consistently pick the innocent suspect as they most closely resemble the culprit in the absence of the guilty suspect. That is, when a witness is presented with a TA line-up, even a 'good' witness (i.e., has a strong memory for the culprit and is explicitly following instructions) will repeatedly pick the innocent suspect until they are given an opportunity to reject that choice during round 3. Another option is to consider information from round 1 (exhaustive binary comparisons) in conjunction with the suspect selection rates from round 3 (final decision). We examined the impact this would have on the *Z*-scores and associated ORs—however, they provided little evidence of improvement. Full details can be found in section 4.2 of the Supporting Information.

Model discussion

Using a theoretical framework guided by the RFC line-up procedure, we developed a model that can be used to assist with the interpretation of a witness' selection behavior and what it could suggest about their memory for the suspect. By comparing how well the model data fit with the data from experiments 1 and 2, we were able to assess, at a group level, children and adults' bias towards selecting the guilty or innocent suspect. Moreover, using the data produced by the model, we demonstrated how ORs can be used to assess individual witness' probable selection bias (or discrimination) for the suspect.

This ability to interpret individual witness behavior is useful. In an experimental paradigm, being able to differentiate between two similar witnesses (like those described earlier) can help us to develop a better understanding of what contextual factors (e.g., exposure quality, age, and attention) can lead to lower levels of guilty suspect selection bias or discrimination. In situations where a witness identifies the suspect but demonstrates a suspect selection bias that is likely to reflect random choosing, this may be used to 'screen out' those who may not be provided sufficient evidence of suspect guilt. Alternatively, even if a witness does not ultimately select a suspect (e.g., rejection or filler), these ORs can be used to interpret witness' bias towards the suspect (i.e., who provides stronger evidence of suspect guilt)—without entirely relying on the line-up decision. This information is particularly useful when considering the desire to learn how much weight an investigator or legal decision-maker should place on the testimony of a particular witness. An important caveat, however, is that these results are based on probabilities and, as such, do not allow for clear discrimination in an applied context between innocent and guilty suspects.

Future directions

Although this initial exploration is promising, there is much work to be performed to further develop the model. As a starting point in assessing the value in such a model, we thought it important to begin using a simple, exploratory approach. That is, we used the model that only varied one of two parameters (how frequently the suspect is selected at each level of selection bias [90% bias—the suspect wins 90% of the time] and kept the filler selection parameter at random or 50%). The advantage of using a simplistic model was that it allowed for a basic examination of whether the model fit the data. Of course, this model was designed as an ideal. In a real-world setting, witnesses may demonstrate a filler selection bias (e.g., through commitment effects, mistaken recognition, and stereotypical 'criminal' appearance)—as opposed to our model that assumed all fillers would be selected at random. Whether the model will fit with data from line-up stimuli that deviate from these parameter assumptions remains an empirical question.

The next natural step is to explore how well the suspect selection bias generalizes to different stimuli. For example, future research should test a model that not only varies how often the suspect is selected but also how often each filler is selected. The ad-hoc similarity ratings between the line-up members could be used to evaluate expected frequency of filler wins in future models. In addition, there is a need to explore whether the suspect selection bias model calibrates with line-up stimuli that has increasing or decreasing suspect-filler similarity. Working to explore these research questions will better establish the utility of the RFC procedure as a tool to discriminate between witnesses that demonstrate bias towards a guilty or an innocent suspect.

GENERAL DISCUSSION

Relying solely on an eyewitness' identification decision often comes with uncertainties about a witness' memory strength and the likelihood of suspect guilt. In the present research, we examined a new, alternative line-up procedure that was designed to provide the same identification decision as traditional line-ups, but also additional information that allows some inferences about a witness' suspect selection bias.

We examined how witnesses from three different age groups (i.e., younger children, older children, and adults) faired using the RFC procedure. Results from experiments 1 and 2 indicate that, for the RFC procedure, age matters. The use of the RFC procedure with young children did not produce favorable results—even increasing inaccurate choosing relative to comparison procedures. However, for children aged 9–11 years and adults, the RFC produced similar levels of identification accuracy as the simultaneous (with wildcard) procedure. Considering the comparable identification performance along with the predictive utility of response pattern information, experiments 1 and 2 results indicate no disadvantage, and in the case of older children, some advantages in using the RFC procedure when compared with the simultaneous procedure. We also demonstrated how a witness' pattern of responding during the RFC procedure can be used to assess a witness' suspect selection bias that, in turn, can be used to understand the likelihood that a witness' responding was based on strong or weak recognition for the suspect. This information may

serve as a tool to learn more about the strength of recognition memory. In addition, this information may be used to screen for reliable witnesses, similar to how the *blank* line-up (i.e., using a known TA line-up prior to a suspect-present line-up; Wells et al., 1998) has been used to identify unreliable witnesses (prone to picking).

Applied implications

Although far too early to recommend applied implementation, if the RFC procedure was to be used in an applied setting, there are several conceivable advantages based on the current data. First, the RFC requires electronic administration,[12] thereby increasing consistency and minimizing administrator influences. Second, we found evidence that the RFC works as well as the simultaneous procedure for both older children (9–11 years) and adult witnesses, allowing for consistent line-up administration for both age groups. Third, and perhaps most importantly, the RFC procedure was designed to provide additional meaningful information. Sauer et al. (2008) discuss the importance of providing a single eyewitness decision for use in an investigation and in court. The RFC procedure provides this categorical line-up decision (i.e., suspect, filler, and rejection)—however, it also provides probabilistic evidence to help investigators and researchers better contextualize that identification.

There are, of course, potential drawbacks of implementing the RFC procedure. Given that the procedure requires 36 decisions (for an eight-person line-up), the extra time and cognitive effort required to complete the task may be considered a barrier. Similarly, interpreting the information (e.g., *Z*-scores and ORs) provided by the RFC procedure in investigations and in the courtroom may require changes to training and test administration protocols.

Theoretical contribution

This research contributes to a greater understanding of eyewitness memory. First, for those over the age of 9 years of age, it demonstrated that eyewitness recognition memory can be assessed using an adjusted forced-choice task. When compared with a simultaneous procedure, the RFC procedure produced similar accuracy. Although, obviously, an increase in accuracy is more desirable, there is a great deal of value in finding new procedures that produce comparable accuracy because it opens the door for further development of these new procedures and new avenues for exploring children's capabilities.

Second, the age difference in performance with the RFC procedure between our younger and older children indicates that developmentally, something important happens around 9 years of age that allows these children to use the RFC procedure more effectively. Younger children's lower discriminatory ability provides evidence that children 8 years and younger may not have the executive functioning ability needed to effectively regulate their responding over a large number of decisions.

[12] Although conducting the RFC procedure in person is possible, it may be vulnerable to human error, and computer administration is much more efficient.

Third, the additional information collected about witnesses while they completed the RFC procedure, including confidence ratings and response patterns, provides insight about the cognitive processes that underlie a witness' decision. In experiment 1, older children's response patterns were predictive of accurate, positive responding—even more so than post-identification confidence ratings. Higher *Z*-scores helped to discriminate between accurate and inaccurate choosers—especially when witness provided ambiguous confidence ratings. However, for adults, *Z*-scores did not predict accuracy better than confidence ratings. As demonstrated by selection bias analyses, the likely reason that *Z*-scores were predictive of accuracy for older children and not adults may result from the ability to adhere to the task instructions. Adults treated the innocent and guilty suspects similarly (i.e., indicated by higher suspect selection bias in TA line-ups), whereas children's choosing associated with the guilty suspect was different from the innocent suspect. Although this finding requires replication, children's discrimination between innocent and guilty suspects with the RFC procedure is worthy of further investigation as a potential source of information about selection bias. Lastly, the model development and analyses indicate that an individual witness' pattern of responding during the RFC procedure can be used to calculate an individual's suspect selection bias, without relying exclusively on their final line-up decision. This, in turn, can help to discriminate between similar-appearing witnesses.

Future research

Although the results of this research are encouraging, further investigation is required. The most important priority for a new procedure is replication. It will be critical to independently replicate the finding of equivalent performance between the RFC and simultaneous (with wildcard) procedure. Likewise, it is important the further test the predictive utility of confidence, latency, and decision patterns of the RFC procedures with different samples to help establish external validity. Second, this research should be extended to explore how RFC performs relative to other line-up procedures—such as the sequential procedure. As children tend to struggle with sequential procedures (e.g., Lindsay et al., 1997), we expect that the RFC procedure will outperform the sequential procedure for this sample. However, with adults, the sequential procedure has been thought to elicit more conservative responding than the simultaneous line-up (Palmer & Brewer, 2012; Steblay et al., 2011) and, as such, how the RFC would perform relative to the sequential procedure remains an empirical question.

Third, there is benefit in exploring other ways to examine what we can learn from witness' decision patterns. After exploring several different options (e.g., max, *M*, variance, and max-*M*/*SD*), we settled on *Z*-scores to index witnesses' suspect selection patterns because it captured the information and variability contained in each witness' suspect selections, relative to his or her filler selections. However, it would be ideal to directly tie response patterns to memory strength, and future research could address this experimentally. For example, one could experimentally manipulate encoding

quality (thereby indirectly manipulating memory strength) and measure the effects on *Z*-score outcomes. If differing patterns were observed in *Z*-scores in response to different encoding qualities, it would provide further support for using patterns of responding to evaluate a witness' selection bias or memory strength for a suspect.

CONCLUSION

The RFC line-up procedure was designed as an alternative method to collect memory evidence from eyewitnesses. We found evidence that the RFC procedure produced similar line-up responding to the simultaneous (with wildcard) procedure for both older children and adult eyewitnesses. This is particularly promising when considering that the RFC procedure, by design, provides additional memory information that can be used to understand the predictive utility of a witness' decision as well as the underlying cognitive processing that may be involved in line-up decisions. These experiments present the first step in demonstrating the utility of using a forced-choice method to examine eyewitness memory.

ACKNOWLEDGEMENTS

This research was supported by a Social Science and Humanities Research Council (SSHRC) scholarship and by an American Psychology-Law Society (AP-LS) grant-in-aid to the first author. This research was also supported by a Natural Sciences and Engineering Research Council Discovery Development Grant to the second author. The authors thank Jessica Llewelyn-Williams, Matthew Pechey, Alyssa Adams, Madison Harvey, Chantalle Fuchs, Ocean Matyjanka, Sarah Ivens, Sara Thomson, Mark Adkins, Amanda Oliver, Odell Tan, Mackenzie Wekerle, Saheba Bajwa, Josh Gonzalas, and Brooke Hoffman for their 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.

REFERENCES

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111. https://doi.org/10.1111/j.0963-7214.2004.01502006.x.

Beaudry, J. L., Lindsay, R. C., Leach, A. M., Mansour, J. K., Bertrand, M. I., & Kalmet, N. (2015). The effect of evidence type, identification accuracy, line-up presentation, and line-up administration on observers' perceptions of eyewitnesses. *Legal and Criminological Psychology*, *20*, 343–364. https://doi.org/10.1111/lcrp.12030.

Bland, J. M., & Altman, D. G. (1995). Comparing two methods of clinical measurement: A personal history. *International Journal of Epidemiology*, *24*(Supplement 1), S7–S14.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*, 700–765. https://doi.org/10.1037/0033-295X.113.4.700.

Bradfield, A. L., & Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. *Law and Human Behavior*, *24*, 581–594. https://doi.org/10.1023/A:1005523129437.

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. https://doi.org/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. https://doi.org/10.1037/1076-898X.12.1.11.

Bruer, K., Fitzgerald, R. J., Therrien, N. M., & Price, H. L. (2015). Lineup member similarity influences the effectiveness of a salient rejection option for eyewitnesses. *Psychiatry Psychology, & Law*, *22*, 124–133. https://doi.org/10.1080/13218719.2014.919688.

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. https://doi.org/10.1007/s11409-012-9091-2.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*, 286–291. https://doi.org/10.3758/BRM.42.1.286.

Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied*, *14*, 139–150. https://doi.org/10.1037/1076-898X.14.2.139.

Deffenbacher, K. A., Leu, J. R., & Brown, E. L. (1981). Memory for faces: Testing method, encoding strategy, and confidence. *The American Journal of Psychology*, 13–26. https://doi.org/10.2307/1422340.

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. https://doi.org/10.1080/13218719.2012.671584.

Dupuis, P. R., & Lindsay, R. C. L. (2007). Radical alternatives to traditional lineups. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology (vol. II): Memory for people*. New York, NY: Lawrence Erlbaum & Associates.

Fitzgerald, R. J., & Price, H. L. (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., Price, H. L., Oriet, C., & Charman, S. D. (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.

Havard, C., & Memon, A. (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.

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. https://doi.org/10.1016/j.jarmac.2013.07.001.

Hood, B. M., Macrae, C. N., Cole-Davies, V., & Dias, M. (2003). Eye remember you: The effects of gaze direction on face recognition in children and adults. *Developmental Science*, *6*, 67–71. https://doi.org/10.1111/1467-7687.00256.

Horry, R., Brewer, N., & Weber, N. (2016). The grain-size lineup: A test of a novel eyewitness identification procedure. *Law and Human Behavior*, *40*, 147–158. https://doi.org/10.1037/lhb0000166.

Humphries, J. E., Holliday, R. E., & Flowe, H. D. (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.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*, 291–306. https://doi.org/10.1037/a0015525.

Jenkins, R., Lavie, N., & Driver, J. (2005). Recognition memory for distractor faces depends on attentional load at exposure. *Psychonomic Bulletin & Review*, *12*, 314–320. https://doi.org/10.3758/BF03196378.

Karageorge, A., & Zajac, R. (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., Brewer, N., & Wells, G. L. (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.

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. https://doi.org/10.1006/jecp.2000.2612.

Lampinen, J. M., Neuschatz, J. S., & Cling, A. D. (2012). *The psychology of eyewitness identification*. New York: Taylor & Francis.

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. https://doi.org/10.1023/A:1024807202926.

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, *70*, 556–564. https://doi.org/10.1037/0021-9010.70.3.556.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research*, *34*, 1703–1721. https://doi.org/10.1016/0042-6989(94)90128-7.

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. https://doi.org/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. https://doi.org/10.1037/a0021034.

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. https://doi.org/10.1007/BF01044534.

Pozzulo, J. D., & Balfour, J. (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., Dempsey, J., & Crescini, C. (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., & Lindsay, R. C. L. (1998). Identification accuracy of children versus adults: A meta-analysis. *Law and Human Behavior*, *22*, 549–570. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/xap0000091.

R. v. Quercia, 1990 ON CA 2595 (November 5, 1990).

Roebers, C. M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology*, *38*, 1052–1067. https://doi.org/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. https://doi.org/10.1016/S0022-0965(03)00076-6.

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

Schneider, W. (1986). The role of conceptual knowledge and metamemory in the development of organizational processes in memory. *Journal of Experimental Child Psychology*, *42*, 218–236. https://doi.org/10.1016/0022-0965(86)90024-X.

Smith, A. M., Lindsay, R. C., & Wells, G. L. (2016). A Bayesian analysis on the (dis)utility of iterative-showup procedures: The moderating impact of prior probabilities. Law and Human Behavior. Advance online publication. https://doi.org/10.1037/lhb0000196

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. https://doi.org/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*, 99–139. https://doi.org/10.1037/a0021650.

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. https://doi.org/10.1037/a0013158.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2014). Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Motivational Science*, *1*, 19–42. https://doi.org/10.1037/2333-8113.1.S.19.

Wells, E. C., & Pozzulo, J. D. (2006). Accuracy of eyewitnesses with a two-culprit crime: Testing a new identification procedure. *Psychology, Crime & Law*, *12*, 417–427. https://doi.org/10.1080/10683160500050666.

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. https://doi.org/10.1037/0021-9010.83.3.360.

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*, 776–784. https://doi.org/10.1037/0033-2909.88.3.776.

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

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*, 155. https://doi.org/10.1037/1076-898X.8.3.155.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*, 603–647. https://doi.org/10.1023/A:1025750605807.

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., & Tannock, R. (1999). Development of inhibitory control across the life span. *Developmental Psychology*, *35*, 205–213. https://doi.org/10.1037/0012-1649.35.1.205.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262–276. https://doi.org/10.1037/a0035940.

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

Zajac, R., & Karageorge, A. (2009). The wildcard: A simple technique for improving children's target-absent lineup performance. *Applied Cognitive Psychology*, *23*, 358–336. https://doi.org/10.1002/acp.1511.

APPENDIX

Appendix A: Z-score Calculation

1. Identify the number of selections of the suspect.
2. Calculate the mean ranking of all lineup members.

$$M_8 = \frac{(\text{Suspect} + \text{Filler 1} + \text{Filler 2} + \text{Filler 3} + \text{Filler 4} + \text{Filler 5} + \text{Filler 6} + \text{Filler 7})}{8}$$

3. Calculate the standard deviation (average distance from the mean) of all lineup members.

$$SD = \sqrt{\frac{\left((M_8\_\text{Suspect 1})^2 + (M_8\_\text{Filler 1})^2 + (M_8\_\text{Filler 2})^2 + (M_8\_\text{Filler 3})^2 + (M_8\_\text{Filler 4})^2 + (M_8\_\text{Filler 5})^2 + (M_8\_\text{Filler 6})^2 + (M_8\_\text{Filler 7})^2\right)}{n-1}}$$

4. Calculate a standardized score (Z-score) by subtracting the mean rankings from the suspect selections and dividing this by the standard deviation.

$$Z - \text{score} = \frac{(\text{Suspect} - M_8\ )}{SD}$$

5. Match the standardized score to Table 8 to assess the likely suspect bias level.

Example

1. Suspect was selected 7 of 7 possible times. Suspect = 7.
2. Mean selection/ranking for all lineup members.
3. Calculate the standard deviation.

Filler 1 = 6, Filler 2 = 5, Filler 3 = 3, Filler 4 = 3, Filler 5 = 2, Filler 6 = 2, Filler 7 = 0

$$M_8 = \frac{(7+6+5+3+3+2+2+0)}{8}$$

$$M_8 = 3.50$$

$$SD = \sqrt{\frac{\left((3.50-7)^2 + (3.50-6)^2 + (3.50-5)^2 + (3.50-3)^2 + (3.50-3)^2 + (3.50-2)^2 + (3.50-2)^2 + (3.50-0)^2\right)}{7}}$$

$$SD = 2.33$$

4. Calculate the Z-score.

$$Z - \text{score} = \frac{(7 - 3.50\ )}{2.33}$$

$$Z - \text{score} = 1.50$$

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.