Flinders
UNIVERSITY

Running Head: POLICE LINEUPS

***American Psychologist* (2019, in press)**

Police line-ups of the future?

Neil Brewer, Nathan Weber, and Nicola Guerin

Flinders University

Author Note

Neil Brewer, Nathan Weber, School of Psychology, Flinders University; Nicola Guerin, now at Centre for Behavioural Research in Cancer, Cancer Council of Victoria.

This research was supported by ARC-DP grants 150101905 and 1093210 awarded to N. Brewer et al. We thank Nicole McCallum and Lizzie Button for assistance with data collection, and D. S. Lindsay for useful discussions on deadline versus no deadline procedures during the early stages of the project. We also acknowledge the conscientious and insightful comments of several reviewers. Some preliminary results from individual studies have been discussed in talks at the conference of the American Psychology-Law Society and the Society for Applied Research in Memory and Cognition.

The data from the experiments are available at https://osf.io/hm2zn/quickfiles

Correspondence concerning this article should be addressed to Neil Brewer, School of Psychology, Flinders University, GPO Box 2100, Adelaide, S. Aust 5001, AUSTRALIA. Email: neil.brewer@flinders.edu.au

Abstract

Problems associated with eyewitness identification decisions have long been highlighted by memory researchers (e.g., Loftus, 1979), with overwhelming evidence that witnesses can err, sometimes with disastrous consequences. Guided by the rationale that witnesses have access to potentially probative memorial information not captured by the traditional categorical lineup responses, an alternative procedure was examined in six experiments with adult (*N*=1669) and child (*N*=273) witnesses. Instead of asking eyewitnesses to identify the offender from the lineup, witnesses rated their confidence in the match between the offender and each lineup member and then variations in the Maximum (*Max*) confidence values assigned (i.e., the highest rated lineup members) were examined. Specifically, we evaluated how well *Max* confidence values predicted suspect guilt or innocence. When suspects (guilty or innocent) in a lineup received the *Max* confidence rating, the probability of guilt increased with the *Max*. When the suspect received a lower rating than the *Max*, they were generally more likely to be innocent. *Max* confidence patterns also predicted guilt where a traditional positive identification would have been unlikely: for example, when the *Max* was low, when the witness gave the *Max* to multiple lineup members, or when a filler received the *Max* but the suspect also received a high rating. The data indicate that witnesses have access to probative memorial information often not captured by the traditional lineup responses of identifying someone or rejecting the lineup. Guidelines for the use and interpretation of this theoretically informed futuristic alternative to existing lineup procedures are provided.

Keywords: eyewitness identification, confidence, probability of guilt, policy reform

Police lineups of the future?

When police are trying to bring a crime perpetrator to justice, they sometimes place a suspect in a lineup and ask a witness to identify the perpetrator. The way lineups are conducted varies considerably across police jurisdictions. The witness may be presented with a live parade or lineup, or a still photo-lineup, or a video of head shots turning from side to side. The number of lineup members can vary (usually from 6 to 12), and they might be presented either simultaneously or sequentially (i.e., one after the other). It is recommended that the lineup should contain only a single suspect among filler photos or individuals that are known to be innocent (see Wells & Turtle, 1986). Many other aspects of lineups vary, but common to all procedures are the possible responses that witnesses may make. Witnesses can positively identify the suspect (who could be innocent or guilty) or one of the lineup fillers (who is known-to-be innocent), or they can *reject* the lineup by indicating that the culprit is not present or that they don't know. The lineup responses made by witnesses in all current procedures are prone to error, sometimes with disastrous consequences. For example, an innocent suspect may be identified and convicted of a serious crime and/or the real perpetrator may not be recognized and thus remain at large in the community. More informative identification evidence could be gained, with less risk of these high stakes errors, by applying lessons learned from recognition memory theory (e.g., Wixted, 2007) and research to the design of an entirely different approach to determining how likely it is that the suspect is the culprit. Here, the efficacy of one approach is demonstrated across six experiments. Regardless of whether the outlined approach ultimately proves to be optimal for accurately evaluating a suspect's guilt, and notwithstanding the established usefulness of extremely high confidence identification decisions for indicating suspects who are very likely to be guilty (see discussion later in this section), it is argued that the criminal justice system's approach to lineups needs to be re-cast into a form that is more compatible with how witness memory works and how it might best be evaluated.

The Problem—And Some Existing Solutions

Problems associated with eyewitness identification decisions have been discussed by memory researchers for many years (e.g., Loftus, 1979). Here, the major problems and the effectiveness of existing solutions to those problems are discussed. Meta-analyses of controlled laboratory studies indicate that perhaps as many as 50% of identification decisions may be inaccurate (Steblay, Dysart, Fulero, & Lindsay, 2003; Steblay, Dysart, & Wells, 2011). And, although there has been considerable controversy about the respective advantages of simultaneous versus sequential presentation of lineups, overall decision error rates are problematically high regardless of the mode of presentation (Steblay et al., 2003, 2011). Documenting the proportion of inaccurate decisions in actual criminal investigations is much more difficult as, unlike in the lab, the identity of the culprit is not known and there are numerous interpretative problems associated with archival data (cf. Horry, Halford, Brewer, Milne, & Bull, 2014). However, the contribution of mistaken identifications to a large proportion of the false convictions overturned through the work of the US Innocence Project (Innocence Project, 2018) is well documented. Also well documented—though, unfortunately, the subject of much less research attention—are the surprisingly high rates (e.g., 35-50%) of lineup rejections (which may or may not be errors) and the non-trivial proportions of filler identifications detected in archival studies (e.g., Horry et al., 2014; Pike, Brace, & Kynan, 2002). Any of these decision errors can have serious consequences: for example, picking an innocent suspect may lead to a false conviction and rejecting a lineup (or even picking a filler) may lead to a dangerous offender avoiding apprehension.

Researchers refined protocols for conducting lineups in order to minimize witness errors, leading to the development and implementation of practices such as ensuring all lineup members match the witness's description and are of similar appearance (Luus & Wells, 1991), warning the witness that the perpetrator may not be present in the array (Steblay, 1997), and ensuring that the lineup administrator is blind to the suspect's identity (Kovera & Evelo, 2017). These and other

practices have been shown to reduce the likelihood of error and certainly can help prevent the serious procedural malpractices that researchers sometimes note when lawyers ask them to examine individual case files. Yet, these advances still leave considerable room for witness error.

Not surprisingly, therefore, eyewitness identification researchers have also expended considerable effort exploring a variety of potential indicators of identification accuracy, including post-identification confidence (e.g., Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013), identification latency (e.g., Brewer, Caon, Todd, & Weber, 2006) and phenomenological reports of decision processes (Palmer, Brewer, McKinnon, & Weber, 2010).

The most well researched and practically feasible of these indicators is the confidence the witness expresses following their identification decision. A number of large-sample studies have shown that, if confidence is recorded immediately after a positive identification decision (i.e., the witness chooses someone from the line-up) and is very high (e.g., 90-100%), the probability of an accurate identification is likely to be high, at least under many of the boundary conditions explored to date (e.g., Brewer & Wells, 2006; Palmer et al., 2013). As confidence falls so too does accuracy, with accuracy much lower at confidence levels below around 80% and extremely low (though certainly not at the floor) when confidence is around 40% or lower (e.g., Brewer & Wells, 2006). Recently, using Mickes' (2015) confidence-accuracy characteristic approach, Wixted and Wells (2017) re-analyzed numerous sets of published confidence-accuracy data, focusing on the accuracy of positive suspect identifications (i.e., excluding any identifications of known-innocent fillers) made with very high confidence. The outcomes of those analyses provide a compelling demonstration that, at the aggregated level, accuracy for extremely confident identifications (i.e., 90-100% confidence) is extremely high when "pristine" lineup conditions prevail (e.g., *inter alia*, a single suspect lineup in which the suspect does not stand out, a confidence statement obtained immediately following the identification). Conversely, low post-identification confidence indicated lower accuracy and poses questions about the reliability of the identification:

"Importantly, a low-confidence ID on an initial test of memory from a lineup signals low accuracy whether or not pristine testing procedures are used. For this reason, low confidence should never be ignored and should instead always raise red flags about the reliability of the ID …" (Wixted & Wells, 2017, p.49). It is also worth noting that, even when pristine conditions apply, very high confidence does not guarantee exceptional accuracy as Wixted and Wells (2017) noted: "… this result serves as a reminder that the determinants of high-confidence accuracy are not fully understood and that more research is needed to identify the conditions under which high-confidence accuracy can be compromised even when fair lineups are used." (p.38).

Wixted and Wells' (2017) analysis of the confidence-accuracy relationship for a traditional lineup suggests that the eyewitness identification problem would be greatly diminished if only decisions made with extremely high confidence at the initial lineup were considered reliable by the criminal justice system. Indeed, Wixted and Wells (2017) highlighted the fact that the low confidence mistaken identifications underpinning many of the wrongful convictions captured by the US Innocence Project (Garrett, 2011) provide independent validation for only trusting extremely confident initial identifications.

Persuading the justice system to rely only on extremely confident initial identifications is clearly one viable approach to the lineup problem. But perhaps eyewitness memory researchers can also offer police, judges, jurors and legislators an alternative to the option of a categorical identification decision (i.e., a positive identification or a rejection) followed immediately by a confidence rating—an alternative that might help answer some of the following questions. For example, what should police, prosecutors and judges make of line-up rejections from witnesses? Examination of data reported in three large-sample lab and field experiments (Brewer & Wells, 2006; Palmer et al., 2013; Sauer, Brewer, Zweck, & Weber, 2010) reveals that 30-35% of lineup rejections were inaccurate: that is, the culprit was actually present in the lineup. It is known that post-identification confidence is not a decisive indicator of the accuracy of non-choosers'

decisions (e.g., Brewer & Wells, 2006; Sporer, Penrod, Read, & Cutler, 1995). Yet, although a rejection may occur because the witness firmly believed the culprit was not in the line-up, it might also occur because the witness was not sufficiently confident to make a positive identification. Or perhaps the witness was confident that the perpetrator was in the line-up but could not decide between two of the line-up members. In both of the latter cases, however, the categorical nature of the identification decision precludes the possibility that the witness may have been able to provide useful memorial information about the culprit's identity.

How should filler picks be interpreted? It is known that filler picks have exculpatory value (Wixted & Wells, 2017). However, might there be circumstances where witnesses can provide useful probative information despite the fact that, given a traditional lineup, they would have picked a filler? Indeed, in human face matching tasks that require participants to find a match for a displayed face in an array of faces, it is not uncommon for a similar filler to be perceived as a better match to memory than the displayed face (e.g., White, Burton, Jenkins, & Kemp, 2014). Thus, it seems quite possible that a witness might consider two lineup members, one of whom is the culprit, as strong matches to memory, but pick the filler because the match seems stronger.

Or, what should police and prosecutors make of identification decisions with less than extremely high confidence (e.g., confidence levels of 70% or 50% or 40%)? Should they assume that the witness's memory is unreliable and thus discard the identification decision? In the three large-sample studies cited above (Brewer & Wells, 2006; Palmer et al., 2013; Sauer et al., 2010) between 63% and 74% of correct identifications were accompanied by confidence ratings below 90%, and between 23% and 40% of correct identifications were accompanied by confidence ratings equal to or less than 60%. Perhaps, such confidence assessments reflect realistic assessments of low quality encoding or a relatively weak encoding-test stimulus match. Nevertheless, they may still offer some probative value. Although some may indicate a witness

who just guessed, perhaps some of those decisions simply reflect the behavior of dispositionally low confident witnesses.

And finally, because post-identification confidence is such a poor indication of the accuracy of children's identification decisions (Keast, Brewer, & Wells, 2007), does this mean that children cannot be relied upon to provide any identity information that offers probative value?

One Alternative Approach

Questions like those just posed highlight the possibility that the limited nature of the identification task's response options means that potentially useful information available in the witnesses' memories may never be accessed. Indeed, we suggest that there is a serious mismatch between the way in which traditional lineup procedures are conducted and interpreted and the current state of knowledge about the memorial processes involved in recognition decisions. Consequently, we argue here for the merits of an alternative approach based on current understanding of the memorial information likely to be available to an eyewitness.

A witness confronted with a traditional lineup either identifies a lineup member (the culprit, an innocent suspect or one of the fillers), whom they believe is the culprit, or rejects the lineup. This categorical decision may be influenced by the quality of the memorial evidence available to the witness as well as by individual difference characteristics and social influences that may shape the criterion the witness sets for making an identification. Although this categorical decision may substantially reduce any ambiguity around a suspect's guilt in cases where an identification is made with extremely high confidence, there are strong theoretical grounds for arguing that, in the case of many identification decisions, it will provide a relatively insensitive evaluation of the match between the witness's memory and the lineup stimuli viewed. Yet, precisely the same theoretical approaches that have led researchers to argue that accuracy should be very high for extremely confident identifications (cf. Wixted & Wells, 2017) suggest an alternative approach to collecting identification evidence that may (a) overcome problems produced by the influence of

non-mnemonic factors on criterion setting for an identification decision and (b) provide probative information that extends beyond that provided by high confidence decisions. As noted by Sauer, Brewer, & Weber (2008), it has long been accepted by recognition memory researchers that signal detection theory perspectives indicate that the match between a memorial representation and a presented stimulus will be indexed by an associated confidence judgment (e.g., Wixted, 2007). Consequently, a previously seen image in a lineup (i.e., the culprit) should generally produce a stronger match to memory, and higher recognition confidence, than hitherto unseen lineup members.

The impetus provided by such perspectives from recognition memory research led researchers (e.g., Brewer, Weber, Wootton, & Lindsay, 2012; Sauer et al., 2008) to propose an alternative to the traditional lineup in which witnesses provide a response for each member of a lineup by indicating their confidence that each is the culprit (i.e., they neither make a decision accompanied by a post-decision confidence judgment nor a rejection). It was argued—with promising initial empirical support—that if witnesses simply indicated their confidence that each lineup member is the culprit then the patterns of confidence judgments for individual lineup members would offer not only useful information about the likelihood that a suspect is the culprit but also more sensitive diagnostic evidence than the traditional lineup decision. Underpinning this approach were three key elements. First, as outlined above, signal detection theories of decision confidence propose that a previously-seen stimulus in a recognition test (e.g., a previously seen offender's face) will generally be associated with a stronger match to memory and higher recognition confidence than unstudied faces (e.g., an innocent suspect or lineup fillers). Second, the information about a witness's memory for the offender provided by the profile of confidence assessments for the memory match with each lineup member will be available no matter whether the witness would have made an identification or rejected a traditional lineup. Third, the potentially rich memorial information accessible to a witness that is provided in the profile of lineup confidence ratings is

both diagnostic of guilt and less likely to be contaminated by other system and witness factors known to affect traditional identification response accuracy.

The Present Study

This study is a major advance on the initial studies that tested confidence profiling procedures and offers a futuristic alternative to the traditional lineup procedures. Novel data analytic procedures reveal a radically different and effective way of using confidence ratings to reframe the eyewitness identification task and interpret the information from the witness, providing a more sensitive index of a suspect's guilt than traditional lineup procedures or, indeed, the initial studies that tested a confidence procedure.[1] Multiple sets of stimuli were used in this study, and the procedure was tested with both adult and child witnesses. Six experiments (five with adult participants, one with children) were conducted in which mock-witnesses assigned confidence estimates to each lineup member and the resulting Maximum (*Max*) confidence patterns were used to predict the probability of suspect guilt (i.e., the likelihood that the suspect is the culprit). As the major data patterns were unaffected by stimuli or experimental manipulations, the presentation of findings focuses on the key patterns from (a) the pooled data from adults in five experiments and (b) the data from child witnesses in one experiment. Pooled adult data provided a substantial dataset that allowed stable estimates for the key patterns. In sum, this approach offers an alternative to existing procedures which removes the dependence on a categorical positive identification, evidence that is so persuasive for jurors (Semmler, Brewer, & Douglass, 2011) and yet so prone to error. We return to this issue in the Discussion section.

The analytic approach is introduced here (and explained in detail in the first part of the Results section). Then, the rationale for the main experimental manipulations is outlined. The focus in analyses was the Maximum (*Max*) confidence rating a witness assigned to one (or multiple) lineup member(s). The confidence scale ranged from 0% (absolutely certain the lineup member is not the culprit) to 100% (absolutely certain the lineup member is the culprit). Therefore, the *Max* value

could lie between 100% (absolutely certain is the culprit) and 0% (absolutely certain not the culprit) and indicated the lineup member(s) who most strongly matched the witness's memory. Each lineup was culprit-present or culprit-absent and, therefore, the *Max* value could have been assigned to a guilty suspect, an innocent suspect, or a known-innocent filler. Whether or not the *Max* value had been assigned to a guilty or innocent suspect, or to a filler, was then examined. If a filler had received the *Max*, the guilty or innocent suspect would have been given a confidence rating lower than the *Max*. If there is only one suspect in the lineup, the critical issue is to diagnose whether the lineup includes a guilty suspect (culprit-present) or an innocent suspect (culprit-absent). This question corresponds to analyzing how well the *Max* predicted the probability that the suspect in the lineup was guilty (i.e., that the lineup was culprit-present, rather than culprit-absent). The probability of guilt could range from 0 to 1.0, with .5 being chance level and 1.0 being 100% likelihood that the suspect in the lineup was guilty. Using this method, the likelihood that the suspect was in the lineup could be evaluated using patterns that sensitively indicated the strength of memory match with the suspect, even if the witness would not have picked the suspect in a traditional lineup (i.e., if the *Max* was too low to support an identification or was given to a filler).

Using this approach, the probability of guilt can be examined for any value of the *Max* confidence (i.e., from as low as 10% to as high as 100%). Further, the probability that the suspect in the lineup is guilty can be evaluated regardless of whether the *Max* is assigned to the suspect or a filler, or whether the *Max* is shared by a suspect and a filler. This is particularly important as the presence of one strong match to memory in the lineup does not rule out the possibility that another lineup member may be an equally strong match. Moreover, if the suspect is one of the *Max* matches and the witness effectively discriminates between multiple *Max* lineup members and the remaining lineup members, the evidence against the suspect may still be valuable for estimating their guilt. Overall, this approach can indicate the probability of guilt when the suspect is a strong

best-match to memory for the offender, or when the suspect is a less strong best-match who would not have been identified from a traditional lineup (perhaps because of poor encoding or a dispositionally uncertain witness). Further, it can indicate likely suspect guilt when the suspect is a strong match but not the best match in the lineup or another lineup member is an equally good match, making it impossible for the witness to decide between them. In other words, the procedure and the analytical approach collectively provide information about whether the suspect in the lineup is guilty that the traditional lineup does not provide.

The first studies that used patterns of *Max* confidence judgments to evaluate suspect guilt each compared a confidence procedure with a traditional lineup. Sauer et al. (2008) asked mock-witnesses to indicate, for each member of a simultaneous lineup, how confident they were that the lineup member was the culprit and used an algorithm to classify the pattern of responses as corresponding to an accurate or inaccurate lineup decision. They then compared the classified accuracy of participants in the confidence ratings procedure with the accuracy of traditional lineup identification decisions (i.e., an identification or a lineup rejection) made by participants who used a traditional simultaneous lineup procedure.

In a slightly different procedure, Brewer et al. (2012) imposed a three-second response deadline on witnesses when making each individual confidence judgment for a lineup member. Lineup members were presented sequentially to ensure each confidence judgment was made within the three-second deadline and accuracy was compared with that in a sequential lineup. It was argued that confidence assessments might more accurately reflect memory strength if potentially biasing metacognitions or heuristics were minimized. Yet, a potential downside of a response deadline is that it limits witnesses' use of 'late-arriving' recollective cues (cf. Rae, Heathcote, Donkin, Averell, & Brown, 2014) that might inform the confidence judgment. Consequently, in the present experiments, a number of experimental conditions were used with adult samples to further investigate the effects of a response deadline on *Max* confidence patterns.

The confidence procedure was also examined with a child sample as child witness errors often result from overly liberal criterion setting for positive identifications (Keast et al., 2007). That is, children respond with similar accuracy to adults when the offender is in the lineup, but are much more likely than adults to select a filler or an innocent suspect from a culprit-absent lineup (Pozzulo & Lindsay, 1998). These patterns suggest that children access useful memorial information about the culprit's identity, despite their greater difficulty than adults in providing accurate identification responses because of their propensity to choose. The confidence procedure potentially provides a way to access that information from children. Bruer, Fitzgerald, Price, and Sauer (2017) recently provided preliminary evidence that classifications of suspect guilt calculated from children's confidence ratings for lineup members were of similar accuracy to children's accuracy in a traditional lineup procedure. However, the analysis approach we propose offers much more nuanced information about suspect guilt.

The General Experimental Approach

In Experiments 1-5, confidence estimates were obtained from adult mock-witnesses who viewed four different mock-crime videos, each followed by 12 sequentially presented lineup members (order of crimes and lineup stimuli randomly assigned).[2] In each experiment *Max* confidence patterns across lineup members were examined for how well they predicted the probability of suspect guilt (i.e., either guilty or innocent suspects from culprit-present or -absent lineups). In each experiment, we manipulated response deadline (deadline, no-deadline lineup) to evaluate the effect of deadlining confidence estimates on witness evaluations of their memory for the offenders, and the usefulness of the confidence procedure. The confidence procedure was evaluated under four conditions predicted to interact with the effect of response deadlines to either undermine or enhance the efficacy of the confidence procedure. The four conditions were: 1) giving witnesses negative feedback about the quality of their offender descriptions; 2) enforcing a confidence response lag in the non-deadlined condition to delay witnesses' confidence estimates;

3) a long retention interval between viewing the crimes and completing the lineups; and 4) dividing witnesses' attention at encoding (Yonelinas, 2002).

Given time to reflect on their confidence judgment (i.e., no deadline), witnesses given negative feedback about the quality of their description of the culprit might doubt their initial rapid assessment (cf. Palmer, Brewer, & Weber, 2010). Similarly, forcing witnesses to withhold their confidence judgments in the no deadline condition was also predicted to leave their confidence estimates open to extraneous metacognitive influences (cf. Brewer, Keast, & Rishworth, 2002). Or, a witness tested after a long retention interval might rapidly assess a strong match to memory for a lineup member, but later adjust their confidence downward because they doubted their memory could be as strong as it seemed after such a delay. In sum, with greater time for reflection, potentially distorting metacognitive influences on confidence may be greater than if a short deadline required them to produce confidence estimates that were more tightly linked to pure memory strength.

Additionally, it seemed possible that the veridicality of witnesses' confidence assessments might be undermined under divided rather than full attention, but with the difference less pronounced under deadline than no deadline conditions. Access to critical recollective cues that provide an important basis for confidence assessments is impaired under divided attention, leaving witnesses to rely more on potentially unreliable familiarity (cf. Palmer, Brewer, McKinnon, & Weber, 2010). As recollective cues are thought to arrive later than familiarity cues (Yonelinas, 2002), a short deadline may deprive witnesses of crucial late-arriving recollected information (cf. Rae et al., 2014). Consequently, confidence judgments may less accurately predict suspect guilt if provided under full attention conditions but made prior to a short deadline. In Experiment 6 the confidence procedure, with no response deadline, was examined with child witnesses and children's accuracy was compared with a traditional sequential lineup.

Method

Participants

All participants in Experiments 1-5 were recruited from the undergraduate and broader university community, and had never participated in an eyewitness identification experiment. There were 1,669 adults (645 male), aged 16 to 69 years (M = 22.59, SD = 7.15)[3]. Experiment 6 participants were recruited from a large suburban school: there were 273 children (128 male) from three grade levels. For those children who recorded their age, age ranged from 10 to 14 years (M = 11.7, SD = 0.9). All child participants completed screening tests to confirm they understood the concepts *present* and *not present* (required for the control condition) and were able to use a 0-100% confidence scale appropriately (10 items). All studies and participant recruitment procedures were approved by the relevant institutional Social and Behavioral Sciences Ethics Committee.

Design

Experiment 1. A 3 (identification procedure: sequential control, deadline confidence, no-deadline confidence) × 2 (pre-identification description feedback: no feedback on description, negative pre-identification feedback on description) × 2 (culprit presence: culprit present, culprit absent) mixed design, with culprit-presence as a within-subjects factor and a 15 minute retention interval between encoding and test. In the confidence procedure conditions, faces were always presented sequentially in all experiments.

Experiment 2. A 2 (identification procedure: deadline confidence, no-deadline confidence) × 2 (pre-identification description feedback: no feedback on description, negative pre-identification feedback on description) × 2 (culprit presence: culprit present, culprit absent) mixed design, with culprit-presence as a within-subjects factor and a 15 minute retention interval between encoding and test. In this partial replication of Experiment 1, for the no-deadline condition confidence response scale buttons presented with each face were visible but could not be activated until 15 s

had elapsed.

Experiment 3. A 2 (identification procedure: deadline confidence, no-deadline confidence) × 2 (culprit presence: culprit present, culprit absent) mixed design, with culprit-presence as a within-subjects factor and a two week retention interval between encoding and test. Activation of confidence buttons was delayed for the no-deadline condition as in Experiment 2.

Experiments 4 and 5. A 2 (identification procedure: deadline confidence, no-deadline confidence) × 2 (attention at encoding: divided, full) × 2 (culprit presence: culprit present, culprit absent) mixed design, with culprit-presence as a within-subjects factor, with a two-week retention interval and delayed activation of confidence buttons for the no-deadline condition.

Experiment 6. A 2 (identification procedure: sequential control, no-deadline confidence) × 2 (culprit presence: culprit present, culprit absent) mixed design with culprit-presence as a within-subjects factor. The retention interval was approximately 3-10 min.

Materials

In each experiment all participants were shown the same four videos of simulated non-violent crimes (15 s to 43 s duration), each involving different events and a different single perpetrator (an offender stealing property from a house; breaking into a car; shoplifting at a supermarket; and stealing a customer's wallet from a café). Previous studies with these stimuli and their associated lineups produced varying identification response patterns and levels of decision accuracy. PCs in individual cubicles presented the stimulus movies and associated lineups; controlled random assignment to conditions, order of movie stimulus presentation and arrangement of lineup stimuli; recorded participant responses; and controlled all experimental manipulations.

The 12 lineup photos for each crime were chest-up front views measuring 8 cm × 6 cm on the monitor. Independent observers provided a free report description of each offender, and lineup fillers and the culprit's replacement in culprit-absent lineups were selected from a pool of photos

that matched the observers' modal descriptions. The culprit replacement (i.e., innocent suspect) was the photo from the pool judged most similar to the offender.[4] To evaluate lineup fairness we analyzed those culprit-absent lineups for which only one lineup member received the *Max* confidence value (i.e., unique *Max*) and found the designated innocent suspects were not too similar or dissimilar to the culprit as they received the *Max* rating in .083, HDI_{95} = [.071, .096], of the lineups (equivalent to chance: 1/12 = .083). A Bayes Factor of 9.77 indicated data were approximately 10 times more consistent with the proportion being at chance levels than they were consistent with being chosen more often than chance.

Procedure

All experiments shared the following procedure, with variations as indicated. Participants were instructed, "You are going to be shown a short film. Pay close attention to it because you will be asked some questions afterwards. When you are ready to watch the film click the "Next" button"; then the movie started. At the end of the first movie, they received the instruction: "Before completing the questions about the film you have just seen, we would like you to watch another film. Once again, please pay close attention." Following four movies, participants completed the lineups after a short retention interval (see above, Experiments 1, 2 and 6) or returned to the laboratory for the lineup phase as close to two weeks later as possible (Experiments 3-5).

In the lineup phase, instructions for viewing the lineup and making responses were displayed on-screen. Participants were instructed which of the crimes the lineup related to before viewing the 12 lineup photos one at a time (i.e., sequential presentation) in randomized order and without any prior information about how many photos they would see. Below each face were 11 on-screen buttons spanning (from left to right) *0% absolutely certain this is not the culprit* to *100% confident that this is the culprit*.

Culprit presence. Culprit presence was counterbalanced, with participants evenly spread across every possible order of two culprit-present and two culprit-absent lineups).

No deadline. After the participant clicked on a button the face disappeared, and the next face and associated confidence scale appeared.

Deadline. Each face was presented for 3 s. After 2 s a buzzer sounded to indicate there was only 1 s left to make a confidence judgment. If a confidence judgment was not made before the deadline, no confidence value was recorded for that lineup member and the lineup face was replaced by the next face.

Retention interval. The retention interval between viewing the videos and the lineups was approximately 5-10 minutes in Experiment 1, 14-15 minutes in Experiment 2, and 3-10 minutes in Experiment 6. In Experiments 3-5, participants completed the lineups as near as possible to two weeks after viewing the videos (range = 7-25 days, M = 14.33, SD = 1.52).

In Experiments 1 (adults) and 6 (children), participants in the control condition were presented with a standard sequential lineup of 12 faces and allowed as much time as they wished to view each photo. If the witness chose a face when presented, the lineup terminated. If the witness rejected a face, it was replaced by the next one. This continued until the participant picked a face or the lineup ended. When the sequential lineup ended, participants indicated their confidence in their final decision by clicking on 1 of 11 on-screen buttons (0%-100%).

Contrary to expectations, other experimental manipulations used in Experiments 1 and 2 (pre-identification description feedback) and Experiments 4 and 5 (attention at encoding) did not interact with the deadline manipulation in their effect on *Max* confidence data patterns. These manipulations are summarized in Supplemental Materials (p. 1).

Results

Data Analytic Approach

Hierarchical Bayesian parameter estimation, essentially a Bayesian version of logistic regression, was used to analyze the data, with *Max* confidence rating as the predictor variable (see description of *Max* predictor variables below) and suspect guilt (culprit-present lineup/culprit-absent lineup) as the outcome (for full model specifications see Supplemental Materials, pp. 2-4.) A distinct predictor variable was used for each of four possible patterns of *Max* confidence ratings gained from two factors: whether the *Max* was given to one or more lineup members (unique/multiple); and whether the suspect was or was not given the *Max* rating (Sus = *Max*/ Sus < *Max*). The four predictors are described here:

(1) Suspect = *Max* Unique: *Max* confidence rating given to the suspect and no other lineup member (*Max* suspect, *Max* unique). (2) Suspect < *Max* Unique: *Max* given to a filler (and suspect given a lower rating) and no other lineup member (*Max* filler, *Max* unique). (3) Suspect = *Max* Multiple: *Max* confidence rating given to the suspect and at least one other lineup member (*Max* suspect, *Max* multiple). (4) Suspect < *Max* Multiple: *Max* given to a filler (and suspect given a lower rating) and at least one other lineup member other than the suspect (*Max* filler, *Max* multiple).

The model was run with each of these four *Max* variables as the predictor, and suspect guilt (culprit-present lineup; culprit-absent lineup) as the outcome, using pooled data from Experiments 1-5 and, separately, using child witness data from Experiment 6. The model for pooled adult data included experiment (1-5), stimulus (i.e., each crime and lineup), retention interval, feedback, attention, and deadline condition as random factors to account for individual differences in responding. Pooling data from Experiments 1-5 provided a larger sample for estimates and allowed us to quantify effects after accounting for variability between experiments (Gelman & Hill, 2007).

As already noted, our outcome variable was the probability of the suspect's guilt.[5] The model estimated differences in this outcome for each *Max* confidence ratings pattern predictor variable (i.e., each of the four calculated *Max* variables described above) between levels of our key factors. Plots of the probability of guilt that was predicted at each level of confidence by each of the four *Max* variables described above are presented (created in R using JAGS; Plummer, 2017). Bayesian 95% highest density intervals (HDI_{95}), similar to frequentist confidence intervals, represent the uncertainty in these estimates. Unlike confidence intervals, 95% HDIs give 95% confidence that the true value of the estimate is within this range and that all values in the range are more plausible than all excluded values. (Full model specifications are provided in the Supplemental Materials, pp. 2-4).

In subsequent sections, this data analytic approach is used to examine: (a) the overall data patterns for Experiments 1-5; (b) the outcomes for the deadline versus no deadline confidence procedure; (c) the impact of retention interval; (d) the impact of suspect-filler confidence rating differences; and (e) response patterns for the sample of child witnesses in Experiment 6.

Overall Data Patterns for Experiments 1-5

First, we provide a qualitative overview of the frequency with which participants provided the various confidence response options. The descriptive statistics for Experiments 1-5 underpinning this overview are reported in Supplemental Materials Table S1 for culprit present, culprit absent and both conditions.

Several features of participants' response patterns emphasize that the traditional categorical lineup responses do not capture the varied nature of the matches that witnesses experience between their memorial image of the culprit and the lineup stimuli. First, although culprit present and absent lineups were presented with equal frequency, the suspect received only about 20% of unique *Max* values (Table S1A). Second, although unique *Max* responses occurred more than twice as often as multiple *Max* responses, multiple *Max* responses still represented more than 30%

of all lineups (Table S1A), with the suspect receiving one of those *Max* values on 45% of the multiple *Max* lineups. In a traditional lineup, such instances might lead to the witness rejecting the lineup or guessing at which of the two stimuli is the culprit when, in fact, the confidence assigned to the suspect may provide probative information.

Two additional (unsurprising) patterns should be noted: (1) The proportion of Suspect = *Max* trials with extremely high confidence (i.e., confidence = 90-100%) was much higher when the *Max* was unique (36.7%) than when shared (7.7%) (Table S1B); (2) The proportion of unique Suspect = *Max* trials with extremely high confidence was much higher for culprit-present (41.0%) than culprit-absent lineups (19.6%) (Table S1B). Two more and important features of witnesses' confidence responses will be described in later sections.

Comparing confidence rating and traditional lineups. A conventional sequential lineup was included in Experiment 1 as a control condition, simply to confirm previous research findings (Brewer et al., 2012; Sauer et al., 2008) showing that the patterns of confidence judgments for individual lineup members provide a stronger indication of suspect guilt than the traditional lineup decision. (See Supplemental Materials Table S2 for control condition confidence data.)

The hierarchical classification algorithm first reported by Sauer et al. (2008) was applied to the Experiment 1 confidence procedure data, with the outcome that overall decision accuracy was much higher than for the sequential control condition (66% vs. 43%, Cohen's $w = 0.23$).

Unique *Max*. To evaluate how well each of the four key patterns of *Max* confidence ratings predicted suspect guilt, we first compared estimates of suspect guilt gained from each of these patterns (i.e., *Max* = Sus Unique; Sus < *Max* Unique; *Max* = Sus Multiple; Sus < *Max* Multiple). Four features of estimates gained for responses when only one lineup member was given the *Max* confidence rating (i.e., *Max* = Sus Unique; Sus < *Max* Unique) should be noted (see plotted estimates of the probability of guilt and associated 95% HDIs, Figure 1, Panel A). First, when suspects received a lower rating than the *Max* confidence rating (Sus < *Max* Unique), the suspect

was more likely to be innocent than guilty (note: one specific exception to this pattern will be highlighted in a later section). Second, when suspects were given the *Max* confidence value (*Max* = Sus Unique), the probability of guilt increased as the *Max* value increased. Specifically, the probability of guilt rose sharply as the *Max* value increased, reaching between .8 and .9 for *Max* values exceeding 75%.[6] Third, when the suspect was given a very low *Ma*x, *Max* values were not informative about the suspect's likely guilt or innocence. Fourth, when the *Ma*x value was given to the suspect and was higher than about 20%, it was more likely that the suspect was guilty than innocent.

It is important to note that unique suspect *Max* values ≥ 30%—that is, *Max* values offering some probative information—comprised 95.7% of all unique *Max* values (Table S1B). Moreover, 61.6% of unique suspect *Max* values offering some probative information (i.e., *Max* ≥ 30%) were below the 90-100% level often classified by eyewitness researchers as indicating extremely high confidence (Table S1B). Thus, the confidence procedure has the potential to provide probative information for substantial proportions of eyewitnesses who would likely not make extremely high confidence identifications.

Shared *Max*. Qualitatively similar patterns were observed for estimates gained from models when the *Max* confidence value was shared (*Max* = Sus Multiple; Sus < *Max* Multiple, see Figure 1, Panel A). When suspects were given a lower confidence rating than a shared *Max* given to two fillers, they were more likely to be innocent than guilty. In contrast, when the suspect was given the *Max* rating along with another lineup member, the probability of suspect guilt increased as the *Max* value increased. *Ma*x values of around 20%-40% that were shared by the suspect and a lineup filler were not informative about guilt or innocence. When these shared values were 0% or 10%, they provided evidence of innocence. Most important, shared *Ma*x values above 40% predicted that the suspect was more likely to be guilty than innocent. Probability of guilt increased with the *Max* confidence value but estimates of suspect guilt from models predicted by a shared

Max confidence rating only reached maximum levels of around .65 and .75 for *Max* values exceeding 75%. That is, *Max* values shared by the suspect and another lineup member less strongly predicted guilt than when the suspect alone was given the *Max* confidence rating in the lineup.

Multiple suspect *Max* values of 50% or higher—that is, multiple *Max* values offering probative indications of guilt—comprised 34.8% of all multiple suspect *Max* values (Table S1B). And, 77.8% of multiple suspect *Max* values of 50% or higher were below the 90-100% level often classified as extremely high confidence (i.e., 90-100%). Thus, by considering multiple *Max* values that include the suspect, the procedure further increases the proportion of eyewitnesses from whom probative information may be obtained.

Response deadline. To evaluate how well the key patterns of *Max* confidence ratings predicted suspect guilt under deadline versus no deadline conditions, estimates of the predicted probability of guilt for each *Max* pattern variable were compared. No effect of the response deadline manipulation (deadline, no deadline; see Supplemental Materials Figure S2) on how well *Max* confidence patterns predicted suspect guilt was evident. The two curves overlap almost completely, providing no indication that deadline procedures enhanced or undermined performance.

Retention interval. In contrast, comparison of estimates predicted by each *Max* pattern variable showed a significant effect of retention interval (immediate, two-week; see Figure 1, Panel B). Specifically, although the probability of suspect guilt predicted by the *Max* confidence pattern was positively associated with the *Max* value for all *Max* patterns regardless of the retention interval, when the *Max* was given to only one lineup member the probability of guilt was higher with the shorter than the longer retention interval when *Max* confidence values were 65% or above (i.e., the 95% HDIs on estimates did not overlap). Retention interval was thus the only variable in Experiments 1-5 that meaningfully affected how well *Max* confidence patterns

predicted suspect guilt. However, when the *Max* confidence rating was shared by the suspect and a lineup filler, there was no evidence of an effect of retention interval on the likelihood that the suspect was guilty (95% HDIs overlapped, see Figure 1).

Suspect-Filler Confidence Rating Differences. In addition to models predicting the likelihood of suspect guilt from the four key *Max* confidence patterns, it was also important to evaluate whether the probative value of the information gained from these *Max* patterns varied depending on how close the ratings for different lineup members were: that is, depending on how close a suspect *Max* rating was to the highest confidence rating given to a lineup filler. Of particular interest was whether there might be cases where the rating for the suspect offers some probative value despite one of the fillers receiving the Max rating.

Earlier (see Footnote 5 and Supplemental Materials Figure S1) it was noted that estimates for the predicted probability of guilt were similar for suspect unique *Max* or the suspect *Max*–the next highest value. When models predicted by *Max* confidence patterns in which the suspect was given a lower confidence rating than the *Max* confidence rating awarded to another lineup member (Suspect < Filler *Max*) were examined, the results were quite surprising.

Figure 2 shows the probability of suspect guilt in relation to the confidence rating assigned to the suspect and to the highest rated filler (*filler-Max*), separately depending on whether the rating for the suspect was equal to or greater than, or less than, the *filler-Max.* High *filler-Max* ratings are denoted by pale blue shading, with progressively lower *filler-Max* ratings denoted by increasingly dark blue shading. The left-hand panel shows that, when the *filler-Max* rating was relatively high, but still lower than the suspect's rating (Sus > *filler-Max*), the probability of suspect guilt was lower than when the difference between the suspect's *Max* rating and the *filler-Max* rating was very large. This, of course, is just another way of representing the suspect *Max*–next highest rating patterns mentioned previously. The surprising finding, however, is shown in Figure 2 (right-hand panel). When the suspect was rated lower than the *filler-Max*, the likelihood

that the suspect was guilty was still higher than the likelihood of innocence if the suspect confidence rating was at least 40% or higher. This indicates that, under some conditions, even though a filler might have been picked in a traditional lineup, the confidence assigned to the suspect can still provide probative information.

Experiment 6 (Child Witnesses)

The frequencies with which child witnesses provided the various response options produced patterns very similar to those already described for adult witnesses. The Experiment 6 descriptive statistics are reported in Supplemental Materials Table S3 for culprit present, culprit absent and both conditions. First, the suspect received only 15.7% of unique *Max* values (Table S3A). Second, although unique *Max* responses occurred about 1.8 times more often than multiple *Max* responses, multiple *Max* responses represented more than 36% of lineups (Table S3A), with the suspect receiving one of those *Max* values on about a third of those lineups. Third, the proportion of Suspect = *Max* trials with extremely high confidence (i.e., confidence = 90-100%) was much higher when the *Max* was unique (51.9%) rather than shared (25.8%) (Supplemental Materials Table S3B).

To enable comparison with a traditional lineup, Experiment 6 included a traditional sequential lineup as a control condition (see Supplemental Materials Table S4 for summary descriptive statistics). Although the choosing rate was high (80.8%), the percentages of accurate suspect identifications and all accurate decisions were very low (10.9% and 13.8%, respectively). Note, however, that the confidence procedure elicited 54 unique Sus = *Max* values and a further 66 multiple Sus = *Max* values (from 536 lineups), while the sequential control condition (139 participants) elicited only 42 suspect identifications (from 556 lineups). Further, 96% (52 of 54) of the unique Sus = *Max* values came from culprit-present lineups versus 83% of accurate suspect identifications (35 of 42) from the sequential control condition.

To determine how well each of the four key patterns of *Max* confidence ratings predicted suspect guilt for child witnesses, a model predicting the probability of suspect guilt was created from confidence ratings for each lineup member using the same *Max* confidence pattern variables as for Experiments 1-5. Due to the much smaller sample in Experiment 6, estimates of suspect guilt predicted by each key *Max* confidence pattern variable are much noisier than for Experiments 1-5. Nevertheless, two features of these patterns of estimates stand out (see Figure 3). First, when suspects received a lower confidence rating than the *Max* and the *Max* was not shared (Sus < *Max* Unique), they were not more likely to be guilty than innocent. Second, when suspects received the unique *Max* (Sus = *Max* Unique), the probability of suspect guilt was very high when the *Max* confidence value was 50% or higher. Unlike the pattern seen with adult witnesses, this pattern was not observed when suspects shared the *Max* value.

Discussion

The propensity for eyewitnesses to err when making identification decisions is well documented. Although various procedural interventions can guard against error arising from administrative malpractices, overall error rates are likely to remain high unless novel approaches to lineups are developed. In this research an alternative approach was outlined and then evaluated in experiments with adult and child witnesses. This approach radically reframes the nature and interpretation of the recognition task. The cornerstone of the approach is that, instead of nominating a single lineup member as the culprit or rejecting the lineup, witnesses simply rate how confident they are that each lineup member is the culprit.

The findings are consistent with the theoretical perspective that adult and child witnesses are likely to have access to memorial information that is informative about the probability of suspect guilt, but is not necessarily detected by a traditional lineup in which the witness either makes a positive identification of a lineup member or rejects the lineup. For adult participants, when the suspect was given the *Max* confidence rating (i.e., the highest confidence rating given to a lineup

member), the probability of suspect guilt rose steeply as that *Max* value rose. And importantly, when suspects were given a lower confidence rating than the *Max*, they were generally (though not always) more likely to be innocent than guilty. It probably comes as little surprise that a suspect who is rated not only as the most plausible lineup member but also as a very strong match to memory is quite likely to be guilty, although the data clearly show that guilt is not guaranteed under those circumstances, especially under conditions likely to weaken memory (e.g., a long retention interval).

The confidence procedure also provided valuable information about likely suspect guilt from child witnesses for whom standard identification responses are problematic because children have difficulty withholding choosing. The data confirm that children within the age range sampled here can access informative memorial information which the traditional identification task often does not access. We caution, however, that the database for children was much more restricted in terms of both sample size and the range of conditions.

Other aspects of the data are perhaps more surprising and provide a powerful demonstration of how accessing more memorial information from the witness than is provided by a categorical decision about the lineup can be so informative about the suspect's likely guilt. Even when the *Max* value was still relatively low (e.g., 30-50%), the suspect was more likely guilty than innocent. Yet, confronted with the traditional lineup decision, a witness with such low confidence that the suspect was the offender may not have made a choice from a culprit-present lineup. For example, a non-choosing witness may have found a relatively good match to memory in the lineup, but have relatively low confidence in this choice because of poor quality of encoding, degradation of the memory trace over the retention interval, or changes in the culprit's appearance over the retention interval. Or a dispositionally unconfident witness with a strong memory match might, if confronted with a traditional lineup, make a positive identification but their confidence level

would not lead to their decision being classified as a highly confident and, hence reliable, identification.

The confidence procedure and the analysis approach used also provided valuable information about the probability of suspect guilt from witnesses who, for other reasons, may not make a choice from a culprit-present lineup. For example, a witness may have find a relatively good match in terms of the suspect but be equally confident about another lineup member, so they decide not to choose. Or a witness may find a very good match in the suspect but may be nearly as confident about another lineup member – and hence does not feel comfortable making a positive decision. These situations present major problems for the standard identification paradigm. Yet overcoming the problem has received very little attention from researchers. As our data show, the confidence procedure can access useful (i.e., diagnostic) memorial information in all of these cases without requiring the witness to decide if they are confident enough to choose or to produce a high confidence identification.

And finally, the confidence procedure provided information about suspect guilt under conditions where witnesses may have picked a known-innocent filler if confronted with a traditional lineup because they were more confident about that filler than about the suspect. Yet, even under those conditions, if the filler attracted high confidence, the suspect was more likely to be guilty than innocent if s(he) also attracted relatively high confidence.

In sum, because the confidence procedure enables the witness to provide a sensitive index of the match between the suspect and memory, it provides useful probabilistic information about suspect guilt that is not provided by many categorical identification decisions. From a theoretical perspective, the various data patterns presented provide a compelling demonstration that the memorial information available to witnesses is continuous in nature. Thus, the data reinforce our introductory claim that traditional police lineups—and, indeed, the way in which police, judges

and jurors tend to conceptualize eyewitness identification evidence—are not aligned with contemporary understanding of the memorial processes underlying such recognition decisions.

Also of theoretical significance is the finding—highlighted in Figure 2 (Panel A) and Figure S1—that probability of suspect guilt increased as the difference between the suspect *Max* and the filler-*Max* increased. This finding is consistent with research with various discrimination and recognition memory paradigms indicating that post-decision confidence is shaped by the relative strength of evidence favoring the chosen alternative over the other available alternatives (e.g., Horry & Brewer, 2016; Van Zandt, 2000; Wixted, Vul, Mickes, & Wilson, 2018). Future research should aim to clarify the optimal diagnostic memory signal.

Type of Confidence Procedure

The only previous research that tested a confidence procedure focused on contrasting group-level identification accuracy by using an algorithm to classify patterns of confidence ratings for lineup members as accurate identifications or rejections. Those studies applied the confidence procedure using either simultaneous lineups (e.g., Sauer et al., 2008) or sequential lineup presentation with confidence judgments required before a deadline had elapsed (Brewer et al., 2012). In this study, deadline and no-deadline sequential lineups were compared to test the possibility that the effectiveness of a response deadline might depend on extraneous conditions. There was no evidence a response deadline restricted access to critical recollective and familiarity information that would guide more veridical confidence judgments. Nor did the absence of a deadline allow witness metacognitions to distort their memory monitoring so as to affect patterns of confidence ratings for lineup members. Whether important differences between such formats might emerge under other conditions is a matter for future research.

Limitations

What, if any, limitations will apply to the generalizability of these findings is unknown. In Experiments 1-5 the culprit-present base rate was 50%. Higher base rates would be expected to

reduce the proportion of *Max* values received by innocent suspects and, in turn, increase the probative value of suspect *Max* values (possibly across the full range of *Max* values). Lower base rates would likely have the opposite effect.

The patterns of reported findings were unaffected by any variations in conditions or stimuli between or within experiments, with the exception of retention interval. Confidence patterns for the two retention intervals remained qualitatively similar, but a high unique suspect *Max* was not as strongly predictive of guilt at the long retention interval as it was with immediate testing. Whether this pattern is sustained with longer retention intervals is unknown. Interestingly, Wixted, Read, and Lindsay (2016) concluded from a re-analysis of a number of retention interval studies using traditional lineups that post-identification confidence was as strongly diagnostic of accuracy at long intervals as it was at short retention intervals. Future research may reveal some critical difference(s) between the confidence procedure and the traditional lineup that explains these apparently discrepant findings. In the absence of a decisive explanation, suffice to say that the data described in our study necessitate the caveat that very high unique *Max* values obtained via the confidence procedure are not as predictive of guilt at long as at short retention intervals.

It is possible that other conditions that contribute to pronounced reductions in memory strength may have a similar effect, although further research is required to confirm this. Similarly, ongoing research should explore how the efficacy of the procedure is affected by factors such as suspect-filler similarity and unfair lineups. Although our analyses detected no significant effect of stimulus materials, increasing suspect-filler similarity beyond the levels captured in these stimuli may produce some increase in multiple *Max* values at the expense of unique *Max* values, perhaps lowering the predictive strength of suspect *Max* values—although recall that overall decision accuracy across all stimuli in the control condition of Experiment 1 was a relatively low 43%, suggesting that the lineup discriminations confronting witnesses were challenging. It is also

possible that a poorly constructed lineup with a standout innocent suspect (and implausible fillers) could contribute to misleading suspect Max values.

Even if future research shows that some variables reduce (or perhaps increase) the predictive power of unique or shared suspect confidence *Max* values, our general position on the application of this approach would remain the same. A major problem with traditional lineups is the tendency to assume that a categorical response such as identifying the suspect or rejecting the lineup denotes guilt or innocence. Under the confidence rating procedure, the eyewitness evidence should be interpreted consistent with estimates of the probability of guilt and their associated HDIs. That is, *Max* values provide nothing more than probabilistic statements that are clearly associated with some measurement error. They should not be seen as offering a definitive statement about guilt or innocence in the way that many seem to interpret positive lineup identifications.

Implications for the Criminal Justice System

Our findings demonstrate that witnesses have access to a rich array of memorial information that can usefully indicate a suspect's likely guilt or innocence. Much of this information does not come to light when using the traditional lineup. Consequently, we are advocating for a world in which the language of correct and mistaken identifications disappears. Instead, if police followed identification procedures similar to those described in these experiments, they would also interpret the confidence information given by witnesses in a similar way as in these analyses. For example, a unique high confidence *Max* value for the suspect would suggest strong grounds for suspicion and further follow-up investigation. A lower unique suspect *Max* value would also provide grounds—albeit weaker—for suspicion. A shared *Max* that included the suspect would also provide grounds for suspicion, and so on. In other words, this type of evidence highlights whether a suspect should remain under suspicion and be viewed as a strong, or perhaps just a very weak, candidate for the culprit. For judges and jurors the same interpretation of confidence evidence should apply, with the evidence considered as just one part of the evidentiary package to be

weighed up. This approach would bring the consideration of identification evidence in line with other scientifically obtained evidence, where the recommendation is not to treat the evidence as proof of guilt (or innocence) but to consider the odds (i.e., probability) of this evidence being observed when the suspect is guilty (vs. innocent) (e.g., Berger, Buckleton, Champod, Evett, & Jackson, 2011)

A decision to follow this approach—or a similar procedure—would require a radical rethink by police, the courts and even eyewitness researchers about how eyewitness judgments should be obtained and interpreted. Many questions would likely arise, some psychological and some legal. As just one example, an obvious psychological issue would be whether jurors could sensibly interpret confidence ratings information. Sauer, Palmer, and Brewer (2017) provide optimistic preliminary data in answer to this question, showing that mock-jurors can draw appropriate conclusions from confidence rating data, although sometimes only with the aid of simple instructions to guide interpretation.

It must be acknowledged, however, that the path to acceptance of this approach and meaningful interpretation of confidence-based evidence could be a challenging one. The traditional police lineup has been used for a very long time and police and prosecutors may rebel against the indecisiveness of the evidence obtained from a confidence procedure as described here. Further, defense lawyers surely would jump at the chance to discredit a witness on the grounds that a *Max* value for the suspect of, say, 50% surely must signal an unreliable memory. They also would likely argue that multiple *Ma*x values for the suspect and a known-innocent filler must suggest the same conclusion. And, of course, knowing that a filler was rated as more likely to be guilty than the suspect would appear to be the death knell for the witness's account. But these examples illustrate why a completely new approach to the interpretation of eyewitness identification evidence is required.

In sum, implementation would obviously be extremely difficult. So, why bother? Reducing the likelihood of people being convicted of crimes they did not commit and of criminals avoiding detection and conviction are, in our view, objectives that warrant not only the careful consideration of psychological researchers but also concerted attempts on their part to translate the relevant psychological science into policy reform.

References

Berger, C. E. H., Buckleton, J., Champod, C., Evett, I. M., & Jackson, G. (2011). Evidence evaluation: A response to the court of appeal judgment in *R v T*. *Science and Justice, 51,* 43-49.

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law and Human Behavior, 30,* 31-50.

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8,* 46-58.

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using deadlined confidence judgments. *Psychological Science, 23,*1208-1214.

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.

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

Denwood, M. J. (2016). runjags: An R Package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC Models in JAGS. *Journal of Statistical Software, 71*.

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical model*s. Cambridge, UK: Cambridge University Press.

Horry, R. & Brewer, N. (2016). How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*, 1615-1634.

Horry, R., Halford, P., Brewer, N., Milne, R., & Bull, R. (2014). Archival analyses of eyewitness identification test outcomes: What can they tell us about eyewitness memory? *Law and Human Behavior, 38*, 94-108.

Innocence Project. (2018). *Innocence Project*. Retrieved August 23, 2018, from http://www.innocenceproject.org/about/index.php

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.

Kovera, M. B., & Evelo, A. J. (2017). The case for double-blind lineup administration. *Psychology, Public Policy, and Law, 23*, 421-437.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distractors for lineups. *Law and Human Behavior, 15*, 43-57.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*, 93–102.

Palmer, M. A., Brewer, N., McKinnon, A. C., & Weber, N. (2010). Phenomenological reports diagnose accuracy of eyewitness identification decisions. *Acta Psychologica, 133,* 137-145.

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

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

Pike, G., Brace, N., & Kynan, S. (2002). *The visual identification of suspects: Procedures and practice*. (Briefing Note 2/02). London: Home Office.

Plummer, M. (2017). *JAGS version 4.3.0 user manual [Computer software manual]*. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/

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

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1226-1243.

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

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347.

Sauer, J. D., Palmer, M. A., & Brewer, N. (2017). Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence. *Law and Human Behavior, 41*, 375-384.

Semmler, C., Brewer, N., & Douglass, A. B. (2011). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons from psychological research* (pp. 185-209). Washington, DC: APA Books.

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

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law & Human Behavior, 21,* 283-297.

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.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showups and lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 27,* 523-540.

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin, 99,* 320-329.

White, D., Burton, A. M., Jenkins, R., & & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20,* 166-173.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152-176.

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5,* 192–203.

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

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81-114.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10–65.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

Footnotes

[1]Previous studies relied on a classification algorithm to determine if the confidence ratings across lineup members indicated likely suspect guilt, and then compared classification accuracy for the confidence procedure group with accuracy on a traditional lineup.

[2]Mansour, Lindsay, and Beaudry (2017) showed that the use of multiple lineup trials for different culprits was not associated with meaningful effects on witness choosing, accuracy or confidence for manipulations of memory strength, disguise type, degree of disguise, and lineup type. Nor were there interactions with variables such as lineup type and memory strength.

[3]Across the five experiments the minimum number of observations per cell was 180. A total of 1,484 participants each completed 4 confidence procedure lineups; 185 participants completed 4 'traditional' sequential lineups in the control condition of Experiment 1.

[4]Our extensive lineup construction experience suggests that using the photo (from a carefully compiled pool of match-description fillers) that appears most similar to the culprit (whether based on a few judges' subjective impressions or numerical similarity ratings) may sometimes produce a 'standout.' Generally, however, inter-individual variability in ratings of different faces is substantial and the innocent suspect turns out (post data collection) to be no more plausible than several other fillers.

[5]Note that our analytic approach focuses on measuring the Bayesian posterior probability of guilt whereas the now quite commonly used ROC approach focuses on assessing discriminability.

[6]The *Max* confidence value is not the only possible index. For example, plotting probability of guilt against the *Max*-next highest confidence value for Experiments 1-5 produced patterns like those shown in Figure 1 (see Supplemental Materials Figure S1).

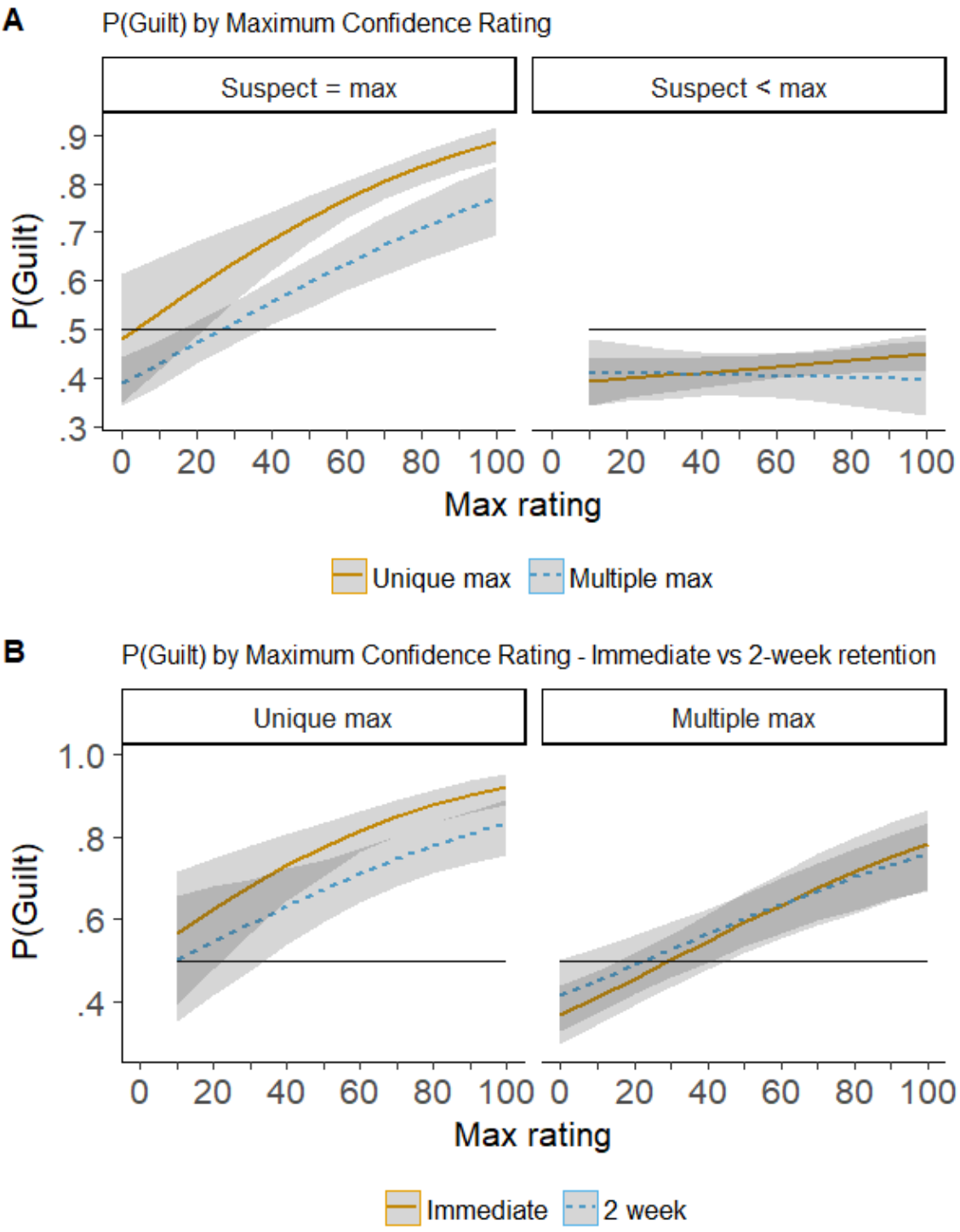

Figure 1. Plots of estimated probability of suspect guilt by *Max* confidence rating assigned, separately for unique and multiple *Max* values. Panel A shows overall data patterns when the Suspect = *Max* and the Suspect < *Max*. Panel B shows patterns for the different retention intervals. The horizontal lines represent the baseline probability of guilt (.5) and the shaded regions indicate 95% HDIs.

Figure 2. Plots of estimated probability of suspect guilt by confidence rating assigned to the suspect and the highest rated filler (*filler-Max*), when the Suspect ≥ *Max* filler rating and Suspect < *Max* filler rating. The black horizontal lines represent the baseline probability of guilt (.5).

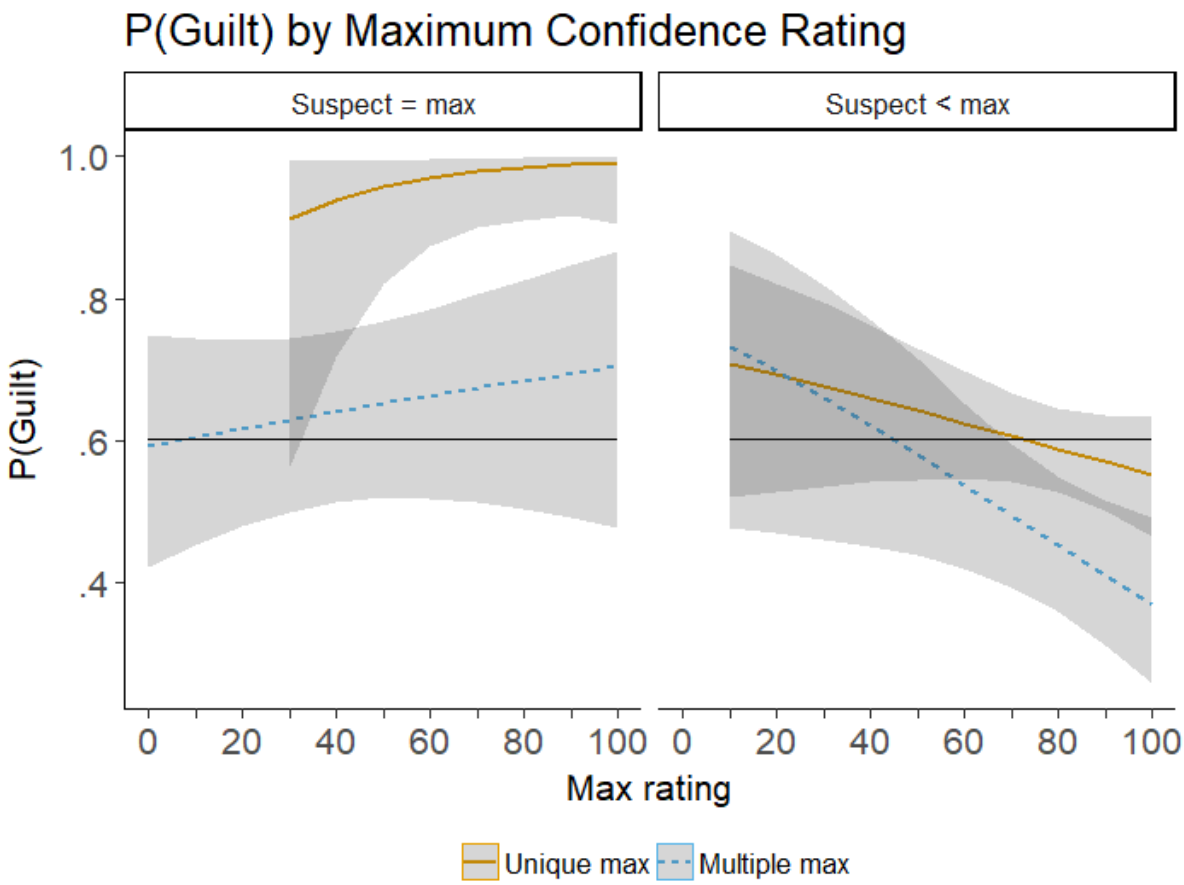

Figure 3. Plots of estimated probability of suspect guilt by *Max* confidence rating assigned, separately for unique and multiple *Max* values when the Suspect = *Max* and the Suspect < *Max* for child witnesses. The horizontal lines represent the baseline probability of guilt (.60, due to the software presenting more culprit-present than -absent lineups) and the shaded regions represent 95% HDIs. There was only one unique suspect = *Max* with a *Max* value below 30 (in that case *Max* = 10), so the plot is constrained to the range of observed values.

Police line-ups of the future?

Supplemental Materials

Experimental Manipulations Not Discussed in the Manuscript

Experiments 1 and 2: Pre-identification description feedback manipulation

- Participants either received no feedback on their description of the culprit or negative feedback on their description.
- No feedback participants were asked to provide the same description as negative feedback participants and, upon completion of the description, were advised to "please wait while the computer prepares the images to be used in the line-ups".
- Negative feedback participants were advised that their descriptions would be matched against a database of ideal descriptions of the offenders and then told, "Your descriptions were VERY POOR. Score: x% match found", where x was a random integer between 10 and 20 (inclusive). Subsequently, participants were informed that description accuracy predicts identification accuracy, thus, "if you were poor at recalling features of other people's faces, then you are, in fact, most likely NOT very skilled at face recognition."

Experiments 4 and 5

- Participants were in either a full or divided attention condition during encoding
- The soundtrack to the stimulus video comprised a series of tones randomized for pitch (high or low) and intervening interval (1 s or 2 s). Participants heard between 101 and 119 tones during the video.
- The divided attention condition required participants to signal the occurrence of low and high pitch tones by pressing keys marked low or high with their left or right index finger, respectively.
- Participants in the full attention condition were told that the soundtrack was relevant to another experiment and asked to ignore the tones.

Full model specifications for Bayesian Mixed Effects

We used hierarchical Bayesian models that are similar to mixed-effects logistic regression to analyze the probability of guilt as a function of the pattern of confidence ratings and relevant experimental factors. In each model, we modeled guilt (vs. innocence) as Bernoulli distributed around a mean probability, μ, that was estimated from a linear combination of predictors via the logistic function. Specifically, to assess the prediction of guilt via *Max* confidence, μ (the probability of guilt) was estimated as a function of an intercept and a coefficient for the slope due to the *Max* confidence value. Separate intercepts and *Max* coefficients were estimated for each of the four possible combinations of two binary properties: whether the suspect's rating was equal to the *Max* versus less than the *Max*; and whether there was a single unique *Max* value versus multiple line-up members received the same *Max* rating. Further, each intercept and slope parameter was allowed to vary as a function of six random effects: retention interval (immediate vs. 2-week), deadline (or not), video (i.e., crime stimulus), experiment, attention (divided vs. full), and feedback (negative vs. none).

Thus, suspect guilt for the *i*th lineup, y^i, was modelled as

$$y^i \sim \text{Bernoulli}(\mu^i)$$

$$\text{logit}(\mu^i) = \beta_0^i + \beta_{Max}^i Max$$

Further, the slope and intercept coefficients, written generically as $\beta_{\cdot}^i$, were estimated as a function of a fixed effect, $b_{\cdot}^{s,u}$, depending on the pattern of confidence ratings, and the random effects, $b_{\cdot}^r, b_{\cdot}^d, b_{\cdot}^v, b_{\cdot}^e, b_{\cdot}^a, b_{\cdot}^f$ reflecting the effects of retention interval, deadline, video, experiment, attention and feedback, respectively. Specifically,

$$\beta_{\cdot}^i = b_{\cdot}^{s,u} + b_{\cdot}^r + b_{\cdot}^d + b_{\cdot}^v + b_{\cdot}^e + b_{\cdot}^a + b_{\cdot}^f$$

The mean of each of the random effect coefficients across levels of that variable was constrained to equal 0 so that the fixed effect coefficient, $b_{\cdot}^{s,u}$, reflects the intercept or slope for that pattern of confidence ratings averaged across variations in experimental conditions.

For both the intercept and slope, the four fixed-effect coefficients, $b_{\cdot}^{s,u}$, were modelled as being drawn from a normal distribution with mean 0 and *SD* 10 (i.e., precision 0.01). The random effect parameters were modelled as drawn from normal distributions with mean 0 and standard deviation estimated from the data. For all standard deviation parameters, $s_{\cdot}^{random}$, we used half-Cauchy priors with location 0 and scale 2.5 (i.e., precision 0.16). Thus the priors for all coefficients were,

$$b_{\cdot}^{s,u} \sim \text{normal}(0, 0.01)$$

$$b_{\cdot}^{random} \sim \text{normal}(0, 1/s_{\cdot}^{random\,2})$$

$$s_{\cdot}^{random} \sim \text{half-Cauchy}(0, 0.16)$$

Guilt was predicted from *Max* minus next using the same model, but with *Max*-next values instead of *Max*.

We used a similar approach to predict guilt from the suspect rating and maximum filler rating (filler-*Max*). Specifically, again guilt was modelled as Bernoulli distributed with probability, μ, estimated via the logistic function from a linear combination of an intercept and the predictors suspect rating, filler-*Max*, and their interaction. Specifically,

$$y \sim \text{Bernoulli}(\mu)$$

$$\text{logit}(\mu) = \beta_0 + \beta_s \times suspect + \beta_f \times filler\text{-}Max + \beta_{s\times f} \times suspect \times filler\text{-}Max$$

As in the previous model the intercept and each slope parameter, generically $\beta_{\cdot}$, were modelled as a fixed-effect, $b_{\cdot}$, with deflections due to each of six random effects (retention interval, deadline, feedback, attention, video, and experiment). We used the same weakly informative priors as for the previous model. Thus,

$$\beta_{\cdot} = b_{\cdot} + b_{\cdot}^{r} + b_{\cdot}^{d} + b_{\cdot}^{f} + b_{\cdot}^{a} + b_{\cdot}^{v} + b_{\cdot}^{e}$$

$$b_{\cdot} \sim \text{normal}(0, 0.01)$$

$$b_{\cdot}^{random} \sim \text{normal}(0, 1/s_{\cdot}^{random\,2})$$

$$s_{\cdot}^{random} \sim \text{half-Cauchy}(0, 0.16)$$

In all analyses, continuous predictors (i.e., *Max*, *max*-next, suspect rating, and filler-*Max*) were scaled to have mean 0 and *SD* 0.5.

Bayesian models predict a probability distribution for estimates, rather than a single point estimate. The probability distributions describe the plausible values for each estimate (i.e., posterior distributions) and the relative credibility of each plausible value. The plausible values can be used to calculate distributions of model predictions (i.e., posterior predictive distributions) that also reflect the credibility of each plausible estimate. Weakly informative priors were set for all model parameters and hyper-parameters. For most complex Bayesian analyses the output distribution has no closed-form solution, so must be estimated via random sampling. Markov Chain Monte Carlo techniques programmed in R (R Core Team, 2017), JAGS (Plummer, 2017) and runjags (Denwood, 2016) were used to generate representative credible values from the joint posterior distribution on the model parameters. Plots of the most credible probability of guilt by maximum confidence value for each condition of interest were created in R using ggplot2 (Wickham, 2009). Analysis of the children's data was identical with the exception that only one random effect (i.e., video) was relevant.

Table S1A

Experiments 1-5 Combined. Number of observations by Max confidence pattern model when the suspect was the culprit or was innocent.

	Unique *Max*		Multiple *Max*	
Culprit	Suspect=*Max*	Suspect<*Max*	Suspect=*Max*	Suspect<*Max*
Present	654	1434	430	438
Absent	163	1807	411	572
Overall	817	3241	841	1010

Table S1B

Experiments 1-5 Combined. Number of observations by confidence level and Max confidence pattern model when the suspect was the culprit or was innocent.

	Confidence												
Culprit	0	10	20	30	40	50	60	70	80	90	100	Total	Possible N
Unique *Max*: Suspect = *Max*													
Present		8	16	26	26	58	57	98	97	100	168	654	2956
Absent		3	8	12	12	27	19	30	20	20	12	163	2953
Overall		11	24	38	38	85	76	128	117	120	180	817	5909
Multiple Confidence Max: Suspect = Max													
Present	142	28	22	26	20	29	32	57	21	22	31	430	2956
Absent	248	28	9	14	11	30	25	21	13	4	8	411	2953
Overall	390	56	31	40	31	59	57	78	34	26	39	841	5909

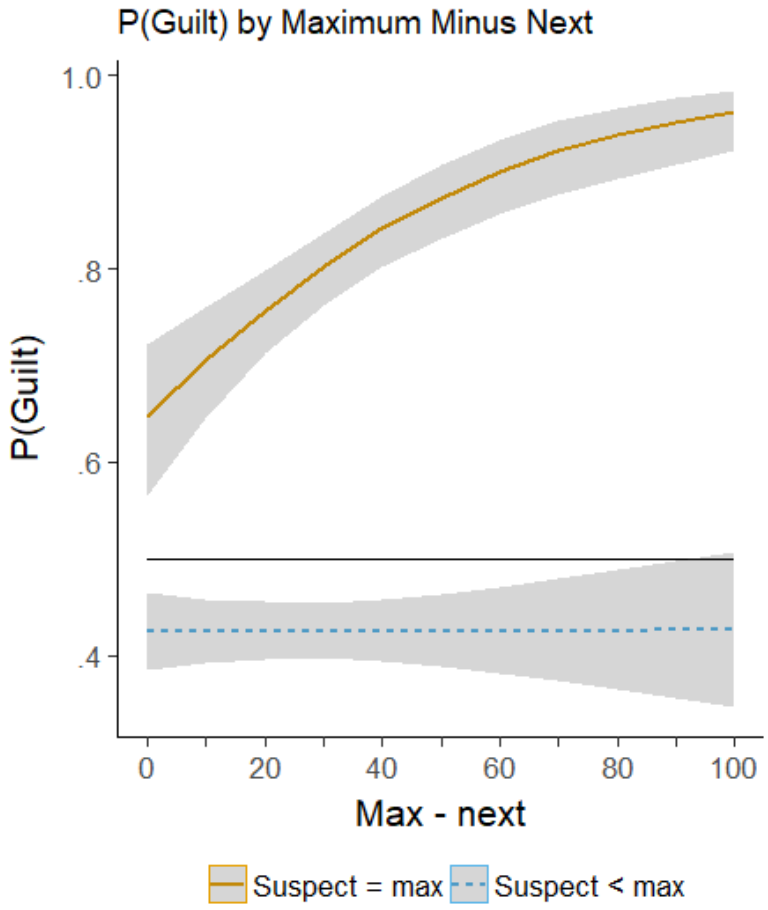

Figure S1. Plots of estimated probability of suspect guilt by *Max*–next highest confidence rating assigned, separately for Suspect = *Max* and Suspect < *Max*. The horizontal line represents the baseline probability of guilt (.5) and the shaded regions represent 95% HDIs.

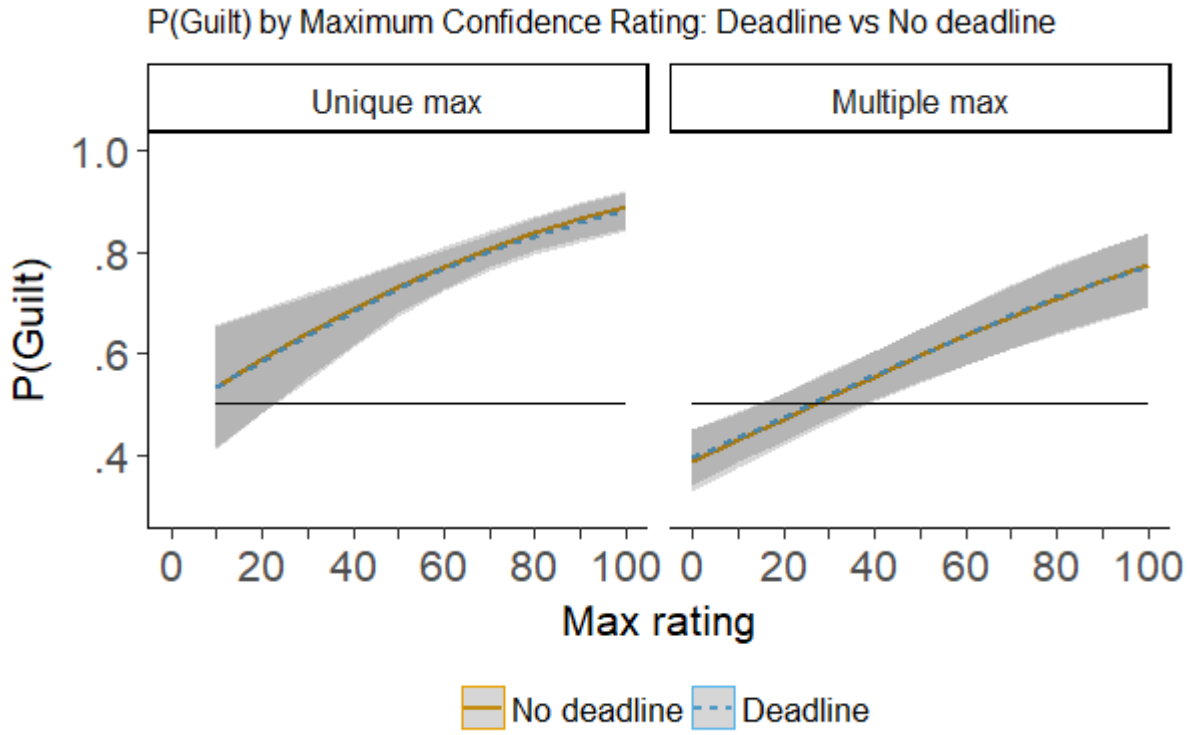

Figure S2. Plots of estimated probability of suspect guilt by *Max* confidence rating assigned, separately for unique and multiple *Max* values for deadline and no deadline conditions for Suspect = *Max*. The horizontal line represents the baseline probability of guilt (.5) and the shaded regions represent 95% HDIs.

Table S2

Experiment 1. Frequency of decision types by confidence and target presence for standard sequential lineups.

	Culprit Present			Culprit Absent		
Confidence	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
0	1	1	9	0	1	17
10	2	1	7	0	2	7
20	1	8	6	0	1	14
30	5	12	9	2	9	13
40	7	15	10	1	18	17
50	18	26	16	0	24	32
60	14	26	10	2	34	23
70	17	29	7	4	28	22
80	21	21	7	0	22	12
90	21	9	12	0	19	19
100	11	2	9	0	3	24

Table S3A

Experiment 6. Number of observations by Max confidence pattern model when the suspect was the culprit or was innocent.[a]

	Unique *Max*		Multiple *Max*	
Culprit presence	Suspect=*Max*	Suspect<*Max*	Suspect=*Max*	Suspect<*Max*
Present	52	170	42	59
Absent	2	119	24	68
Overall	54	289	66	127

[a]134 participants were allocated to the confidence procedure condition.

Table S3B

Experiment 6. Number of observations by confidence level and Max confidence pattern model when the suspect was the culprit or was innocent.

	Confidence												
Culprit	0	10	20	30	40	50	60	70	80	90	100	Total	Possible N
Unique *Max*: Suspect = *Max*													
Present		1		4	1	2	4	6	6	13	15	52	323
Absent[a]					1			1				2	213
Overall		1		4	2	2	4	7	6	13	15	54	536
Multiple Confidence *Max*: Suspect = Max													
Present	17	2		2	2	3	1	2	1	2	10	42	323
Absent	12	2	1	1				1	2		5	24	213
Overall	29	4	1	3	2	3	1	3	3	2	15	66	536

[a]Note that software presented more culprit-present than -absent lineups.

Table S4

Experiment 6. The response patterns for the sequential lineup control condition.[a]

Culprit presence	Suspect ID	Filler ID	No ID
Present	35	220	65
Absent	7	187	42
Overall	42	407	107

[a]139 participants were allocated to the sequential control condition.