Laboratory-based mock crime studies have often been interpreted to mean that (i) eyewitness confidence in an identification made from a lineup is a weak indicator of accuracy and (ii) sequential lineups are diagnostically superior to traditional simultaneous lineups. Largely as a result, juries are increasingly encouraged to disregard eyewitness confidence, and up to 30% of law enforcement agencies in the United States have adopted the sequential procedure. We conducted a field study of actual eyewitnesses who were assigned to simultaneous or sequential photo lineups in the Houston Police Department over a 1-y period. Identifications were made using a three-point confidence scale, and a signal detection model was used to analyze and interpret the results. Our findings suggest that (i) confidence in an eyewitness identification from a fair lineup is a highly reliable indicator of accuracy and (ii) if there is any difference in diagnostic accuracy between the two lineup formats, it likely favors the simultaneous procedure.

Estimating the reliability of eyewitness identifications from police lineups

John T. Wixted\textsuperscript{a,1}, Laura Mickes\textsuperscript{b}, John C. Dunn\textsuperscript{c}, Steven E. Clark\textsuperscript{d}, and William Wells\textsuperscript{e}

\textsuperscript{a}Department of Psychology, University of California, San Diego, La Jolla, CA 92093; \textsuperscript{b}Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, United Kingdom; \textsuperscript{c}School of Psychology, The University of Adelaide, Adelaide, SA 5005, Australia; \textsuperscript{d}Department of Psychology, University of California, Riverside, CA 92521; and \textsuperscript{e}Department of Criminal Justice and Criminology, Sam Houston State University, Huntsville, TX 77341

Edited by Daniel L. Schacter, Harvard University, Cambridge, MA, and approved November 9, 2015 (received for review August 23, 2015)

Eyewitnesses to a crime are often called upon by police investigators to identify a suspected perpetrator from a lineup. A traditional police lineup in the United States consists of the simultaneous presentation of six people, one of whom is the suspect (who is either guilty or innocent) and five of whom are fillers who resemble the suspect but who are known to be innocent. Live lineups were once the norm, but, nowadays, photo lineups are much more commonly used (1). When presented with a photo lineup, an eyewitness can identify someone—or the suspect (a suspect ID) or one of the fillers (a filler ID)—or can reject the lineup (no ID). A filler ID is a known error that does not merit the individual identified, but a suspect ID (including a misidentification of an innocent suspect) does. According to the Innocence Project, eyewitness misidentification is the single greatest cause of wrongful convictions in the United States, having played a role in over 70% of the 333 wrongful convictions that have been overturned by DNA evidence since 1989 (2).

In an effort to reduce eyewitness misidentifications, several reforms based largely on the results of mock crime studies have been proposed. In a typical mock crime study, participants become witnesses to a staged crime (e.g., a purse snatching) and then later attempt to identify the perpetrator from a target-present lineup (containing a photo of the perpetrator) or a target-absent lineup (in which the photo of the perpetrator is replaced by a photo of the “innocent suspect”). The results of mock crime studies have often been interpreted to mean that (i) eyewitness confidence is an unreliable indicator of accuracy (3, 4) and (ii) suspect ID accuracy is enhanced—and the risk to innocent suspects is reduced—when the lineup members are presented sequentially (i.e., one at a time) rather than simultaneously (5–7). In light of such findings, the state of New Jersey recently adopted expanded jury instructions stating that eyewitness confidence is a generally unreliable indicator of accuracy (8). In addition, up to 30% of law enforcement agencies in the United States that use photo lineups have switched to using the sequential procedure (1).

The idea that eyewitness memory is generally unreliable has undergone revision in recent years, as has the notion that sequential lineups are diagnostically superior to simultaneous lineups. With regard to the reliability of eyewitness identifications, recent mock crime studies using a calibration approach have provided strong evidence that confidence in a suspect ID from a photo lineup can be a highly reliable indicator of accuracy (e.g., refs. 9–12). Whether this is true of real eyewitnesses remains unknown and is the first focus of a new police department field investigation that we report here. Previous police department field studies of eyewitness confidence are rare. Those that have been performed found that confident eyewitnesses were more accurate than less confident eyewitnesses (13, 14). However, the investigating officer who administered the lineup knew who the suspect was, raising the possibility that this effect merely reflected administrator influence.

With regard to lineup format (simultaneous vs. sequential lineups), recent mock crime studies using receiver operating characteristic (ROC) analysis (15–17) have generally found that simultaneous lineups are, if anything, diagnostically superior to sequential lineups (18–21). Similarly, in a recent police department field study comparing the two lineup formats, expert ratings of evidence against identified suspects favored the simultaneous procedure (22). However, a different analysis based on filler ID rates from the same field study was interpreted as supporting the sequential procedure (23). Determining which lineup format is diagnostically superior is the second focus of our investigation.

Our field study was conducted in the Robbery Division of the Houston Police Department (24). We focus here on a subset of criminal investigations initiated by the department in 2013 that (i) used photo lineups pseudorandomly assigned to simultaneous (\( n = 187 \)) or sequential (\( n = 161 \)) formats, (ii) were administered by an investigator who was blind to the identity of the suspect, and (iii) involved suspects who were strangers to the eyewitnesses.

Significance

In contrast to prior research, recent studies of simulated crimes have reported that (i) eyewitness confidence can be a strong indicator of accuracy and (ii) traditional simultaneous lineups may be diagnostically superior to sequential lineups. The significance of our study is that these issues were investigated using actual eyewitnesses to a crime. Recent laboratory trends were confirmed: Eyewitness confidence was strongly related to accuracy, and sequential lineups were, if anything, diagnostically superior to sequential lineups. These results suggest that recent reforms in the legal system, which were based on the results of older research, may need to be reevaluated.

Author contributions: S.E.C. and W.W. designed research; W.W. performed research; J.T.W., L.M., J.C.D., and W.W. analyzed data; and J.T.W., L.M., J.C.D., S.E.C., and W.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. Email: jwixted@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516814112/-/DCSupplemental.
Eyewitnesses who made suspect IDs or filler IDs from these lineups were asked to supply a confidence rating using a three-point scale (high, medium, or low confidence). These lineups are of particular interest because they correspond to the “double blind” lineup administration procedure that was recently recommended by a committee of the National Academy of Sciences on eyewitness identification (25). In SI Results, we present a similarly detailed and largely convergent analysis of 194 simultaneous and 175 sequential lineups from a “blinded” condition in which the lineup administrator knew the identity of the suspect but was blind to the position of the suspect in the lineup. In analyzing the results, we not only report empirical trends but also offer a quantitative theoretical interpretation of the data by drawing upon standard models of recognition memory.

Results

Lineup Fairness. Lineup fairness was examined for a random sample of 30 photo lineups from the blind condition (15 simultaneous and 15 sequential). This analysis assessed the degree to which the suspect stood out by providing the selected photo lineups to 49 mock witnesses and asking them to try to identify the suspect based only on the suspect’s physical description. In a fair, six-person lineup, the suspect should be identified by a mock witness only 1/6 (0.17) of the time. The mean proportion of suspect IDs made by the mock witnesses (0.18) did not differ significantly from the expected value for a fair lineup, t(29) = 0.76.

Confidence in Suspect IDs and Filler IDs. We next analyzed eyewitness identification collapsed across lineup format (i.e., simultaneous and sequential data combined; see Table S1). Suspect IDs, filler IDs, and no IDs (Fig. 1A) occurred with approximately equal frequency. The relatively high frequency of filler IDs (which are IDs of known innocents) could be interpreted to mean that eyewitness memory is unreliable (7), but it is important to keep in mind that there are 5 times as many fillers as suspects in a lineup. Moreover, most filler IDs were made with low confidence, whereas most suspect IDs were made with high confidence (Fig. 1B). In other words, the proportion of IDs that were suspect IDs increased markedly with confidence (Fig. 1C). This pattern of results immediately suggests a strong relationship between confidence and accuracy.

Corroborating Evidence. For each lineup, the investigating officer indicated whether or not there was independent corroborating evidence of suspect guilt (see Table S2). The proportion of lineups associated with such evidence was higher for lineups involving suspect IDs (97 out of 114) than lineups involving no IDs (67 out of 130), $\chi^2(1) = 31.02$, $P < 0.0001$, suggesting that suspects identified by an eyewitness were more likely to be guilty than suspects who were not identified by an eyewitness. In addition, for the suspect IDs, the proportion of cases with corroborating evidence of guilt increased as confidence in the ID increased (Fig. 1D). The existence of corroborating evidence was a subjective interpretation made by the investigating officer. However, the results were virtually unchanged when a five-member research team reviewed and recoded the existence of corroborating evidence in a few instances where a majority of the team members disagreed with what the investigating officer counted as independent evidence (see SI Results, Recoded Corroborating Evidence).

Although the data in Fig. 1C imply that suspect ID accuracy increased with confidence, the dependent measure in that figure, namely, suspect IDs/(suspect IDs + filler IDs), includes all suspect IDs (guilty suspect IDs + innocent suspect IDs). A measure of greater interest to the legal system is suspect ID accuracy: guilty suspect IDs/(guilty suspect IDs + innocent suspect IDs). This measure is of greater interest because, as a general rule, only suspects who are identified from a lineup are placed at risk of prosecution. Suspect ID accuracy cannot be directly computed in a police department field study because it is not known which identified suspects are guilty and which are innocent, but it can be estimated using a model of recognition memory.

Two traditional and often competing approaches to modeling recognition memory are the “high-threshold” modeling approach and the signal detection modeling approach (26). Our goal here is not to determine which approach is more viable for modeling eyewitness identification performance but is to instead show that, despite being based on completely different assumptions, both approaches provide similar interpretations of the Houston field data. We begin by using a simple version of the high-threshold model to interpret the data and then provide a more detailed interpretation of the same data using a signal detection model.

High-Threshold Estimates of Suspect ID Accuracy. A virtue of the high-threshold approach is that it provides an algebraic estimate of suspect ID accuracy. According to this model, the witnesses presented with a target-present lineup, some proportion of them, $p$, will recognize and correctly identify the perpetrator. Of the remaining proportion of those witnesses, $1−p$, some proportion of them, $g$, will make a random identification from the lineup despite not recognizing the perpetrator. For a fair, six-member lineup, these witnesses will, by chance, correctly identify the perpetrator 1/6 of the time, and they will instead identify a filler 5/6 of the time. Thus, the probability of a correct suspect ID from a target-present lineup is equal to the probability that a witness recognizes the perpetrator, $p$, plus the probability that a witness who does not recognize the perpetrator makes a lucky guess, $(1−p) · g · (1/6)$. Multiplying the sum of these probabilities by the number of target-present lineups, $n_{TP}$, yields the predicted number of suspect IDs from target present lineups, $n_{STP}$,

$$n_{STP} = n_{TP} · [p + (1−p) · g · (1/6)].$$

The probability of a filler ID from a target-present lineup is equal to the probability that a witness who does not recognize the perpetrator makes a guess that lands on a filler, $(1−p) · g · (5/6)$. Thus, the number of filler IDs from target-present lineups, $n_{FTP}$, is
\[ n_{FP} = n_{FP} \cdot [(1 - p) \cdot g \cdot (5/6)]. \]  

For witnesses presented with target-absent lineups, the state of recognition theoretically does not occur because the guilty suspect is not there, so innocent suspect IDs and filler IDs are only made by witnesses who make a random guess. As indicated above, a random guess occurs with probability \( g \). Thus, the probability of an incorrect (i.e., innocent) suspect ID from a fair target-absent lineup is \( g \cdot (1/6) \), and the probability of a filler ID from a fair target-absent lineup is \( g \cdot (5/6) \). Multiplying these probabilities by the number of target-absent lineups, \( n_{TA} \), yields the predicted number of suspect IDs and filler IDs from target-absent lineups,

\[ n_{STP} = n_{TA} \cdot g \cdot (1/6) \]  

\[ n_{FTA} = n_{TA} \cdot g \cdot (5/6). \]

These equations underscore the important fact that, for fair lineups, incorrect suspect IDs should be relatively rare compared with incorrect filler IDs.

In a study of real police lineups, the information that is known consists of the number of lineups administered, \( N \), the number of suspect IDs, \( S \), the number of filler IDs, \( F \), and the number of no IDs. In terms of the model, \( S \) is equal to sum of suspect IDs from target-present and target-absent lineups (Eq. 1 + Eq. 3) and \( F \) is equal to sum of filler IDs from target-present and target-absent lineups (Eq. 2 + Eq. 4),

\[ S = n_{TP} [p + (1 - p) \cdot g \cdot (1/6)] + n_{TA} \cdot g \cdot (1/6) \]

\[ F = n_{TP} [(1 - p) \cdot g \cdot (5/6)] + n_{TA} \cdot g \cdot (5/6). \]

If, for the sake of simplicity, we assume equal base rates such that \( n_{TP} = n_{TA} = n \), where \( n = N/2 \), then we can algebraically solve for \( g \) and \( p \) (SI Results, High-Threshold Model), which yields

\[ g = (6 \cdot F)/(10 \cdot n - 5 \cdot S + F) \]  

\[ p = (5 \cdot S - F)/(5 \cdot n). \]

Note that, using Eqs. 5 and 6, \( p \) and \( g \) can be directly computed from the data because they are both a function of known values (\( S, F, \) and \( n \)). With \( p \) and \( g \) in hand, Eqs. 1 and 3 can now be used to estimate \( n_{STP} \) and \( n_{STA} \), which can then be used to compute suspect ID accuracy, \( S_{acc} \).

\[ S_{acc} = n_{STP}/(n_{STP} + n_{STA}). \]

\( S_{acc} \) is the measure of interest. As an example, there were 348 blind lineups \((n = 348)\). Therefore, assuming equal base rates, \( n = N/2 = 174 \). There were 114 suspect IDs \((S = 114)\) and 104 filler IDs \((F = 104)\). According to Eqs. 5 and 6, \( g = 0.49 \) and \( p = 0.54 \). Using these parameters, Eqs. 1 and 3 indicate that \( n_{STP} = 99.8 \) and \( n_{STA} = 14.2 \), so overall suspect ID accuracy (Eq. 7) comes to 99.8/\(99.8 + 14.2 = 0.88 \) (i.e., 88% correct).

A similar high-threshold model can be used to predict suspect ID accuracy separately for each level of confidence by following the same computational steps as before, but, this time, using the number of suspect IDs and filler IDs made with a specific level of confidence in place of the overall \( S \) and \( F \) values. Although the computational steps are exactly the same, the implied underlying model now involves additional parameters that allow for different levels of confidence to be expressed when the witness is in the detect state or in the guessing state (see SI Results, High-Threshold Model). This version of the model has as many parameters as there are degrees of freedom in the data, so it cannot be independently validated (e.g., using a goodness-of-fit test).

Nevertheless, the model can still be used to directly estimate suspect ID accuracy separately for each level of confidence using the same computational steps that were used above for overall suspect IDs and filler IDs. When confidence-specific suspect ID and filler ID values are used, the estimated suspect ID accuracy scores come to 0.97, 0.87, and 0.64 for high-, medium-, and low-confidence IDs, respectively. In addition, when this theoretical analysis is performed separately on the data from the blind simultaneous and blind sequential conditions collapsed across confidence, \( p \) (the probability of successfully identifying the perpetrator from a target-present lineup) is 0.62 for simultaneous lineups and 0.43 for sequential for sequential. The significance of these apparent trends cannot be tested, because the model is saturated. We turn now to a more detailed model-based analysis using signal detection theory. This model has fewer free parameters, so its interpretation of the data can be statistically evaluated. We first fit the model to data from an experimentally controlled study (as a validation test) and then fit the model to the data from the Houston field study.

**Signal Detection Estimates of Suspect ID Accuracy.** In the context of eyewitness memory, the standard unequal variance signal detection model (Fig. 2) (26–28) specifies how memory strength is distributed across guilty suspects (targets) vs. innocent suspects and fillers (lures). Before applying this model to the Houston field data, we first tested its validity in the context of eyewitness identification by evaluating its performance in relation to data recently collected as part of a large-scale \((n = 908)\) investigation into the relationship between confidence and accuracy under naturalistic conditions (similar to a mock crime study). In this study, the experimenters approached participants in parks and shopping malls and asked them to view a target person (11). Participant memory for the target (the “guilty suspect”) was subsequently tested using an eight-person simultaneous photo lineup, with half of the participants being tested with a target-present lineup and the other half with a target-absent lineup. Thus, in this study, it was known whether a suspect ID was correct or incorrect.
The observed identification decisions (Fig. 3A) can be collapsed across target-present and target-absent lineups (Fig. 3B), as if this study were a police department field study with unknown lineup type, thereby allowing a comparison with the analogous Houston Police Department field data (Fig. 1A). When the data are broken down by confidence (Fig. 3C), the trends are similar to the trends observed in the Houston field data (Fig. 1B).

How well does the signal detection model (Fig. 2) characterize the experimentally controlled field data (Fig. 3C)? Ordinarily, the parameters of the model would be adjusted to minimize the χ² goodness of statistic between the predicted target-present and target-absent data vs. the observed target-present and target-absent data in Fig. 3A (see SI Results, Signal Detection Model Fits). However, if these data had come from a police department field study, that kind of evaluation would not be possible because it would not be known which lineups contain a guilty suspect (target-present) and which contain an innocent suspect (target-absent).

We therefore fit the signal detection model to the experimentally controlled field data as if those data had come from a police department field study. For each iteration of the fit, the model (Fig. 2) generated simulated predicted target-present and target-absent data, which were then collapsed across lineup type to yield predicted suspect IDs and filler IDs (for three levels of confidence in each case), plus predicted no IDs for that iteration. The collapsed predicted values were then compared with the collapsed observed values by computing a χ² goodness-of-fit statistic. The model assumed equal base rates for target-present and target-absent lineups, which is known to be true of these data (11), and the model parameters were adjusted to minimize the predicted vs. observed χ² statistic, yielding the final predicted values in Fig. 3D. An equal variance model turned out to be adequate (i.e., σtarget did not differ significantly from 1; thus σtarget = σlineup). When the observed data (Fig. 3C) and predicted data (Fig. 3D) were used to compute the observed and predicted proportion of IDs that were suspect IDs, the two functions were nearly identical (Fig. 3E).

Using the experimentally controlled field data (11), we can now ask how the observed trend in Fig. 3E based on data collapsed across target-present and target-absent lineups relates to suspect ID accuracy (the measure of primary interest), which, unlike in a police department field study, can be directly computed after disaggregating the target-present and target-absent data. The actual disaggregated suspect ID accuracy data from this study reflect highly reliable eyewitness ID performance (Fig. 3F). Remarkably, the model accurately predicted those data (Fig. 3F) despite having only been fit to the collapsed (real-world-like) data (Fig. 3C and Table S3).

Having established that the signal detection model can recover suspect ID accuracy from collapsed data, we next fit the model to the Houston Police Department field data (i.e., to the data shown in Fig. 1B), for which it is impossible to separate target-present and target-absent lineups. Initially, we made the assumption that the base rate of guilty suspects (i.e., the proportion of target-present lineups) in these real-world data was 50%. The validity of this assumption is unknown, so we repeated the model-fitting exercise assuming a 25% base rate and, then, a 75% base rate. For all of these fits, we allowed σtarget and σlineup to differ. The model
was fit to the simultaneous and sequential data separately and also to the data combined across lineup format (Tables S4 and S5; also see Tables S6 and S7). Critically, we can use the best-fitting model to estimate the accuracy of suspect IDs in the Houston data, just as we did for the data shown in Fig. 3F. Because the suspect ID accuracy estimates were very similar for the two lineup formats, we present the results of the fit to the data combined across lineup format.

Fig. 4A shows the estimated suspect ID accuracy ($S_{cr}$) for the Houston field data—that is, it shows estimated values of $n_{STP}$ ($\eta_{STP}$ + $n_{STA}$) as a function of confidence for each of the three base rates considered. These data represent the predicted posterior probability of guilt associated with suspect IDs made with low, medium, or high confidence. The estimates for high-confidence suspect IDs remain very accurate regardless of the base rate, whereas the estimated accuracy of low-confidence suspect IDs is always lower but varies considerably depending on the base rate of guilty suspects in police lineups.

**A Model-Based Estimate of the Target-Present Base Rate.** Based on the results of the model fit to the experimentally controlled field data (11), we next made the assumption that an equal variance model ($\sigma_{Target} = \sigma_{Lum}$) also applies to the Houston field data. Removing the unequal variance parameter made it possible to add a base rate parameter ($\mu_{Target}$) to the model to obtain a principled estimate of the real-world base rate of target-present lineups (see SI Results, Signal Detection Model Fits). Again using the experimentally controlled field data (11), we first verified that when target-present and target-absent data are combined in varying proportions and then fit with the equal variance signal detection model, the base rate of target-present lineups can be accurately recovered (Fig. S1). We then fit the equal variance model (including the base rate parameter) to the Houston Police Department field data, and the estimated base rate of target-present lineups came to 0.35 for both simultaneous and sequential data separately and also when the reduced recoded data set was analyzed (excluding the 65 witnesses discussed above), the differences on the three questionnaire measures, if they were true and had any effect, would have worked against the sequential procedure. When these 65 witnesses were excluded from the analysis, the proportions of identified suspects from simultaneous lineups ($n = 50$) and sequential lineups ($n = 38$) rated as having independent evidence of guilt against them were virtually unchanged ($SIM = 0.920$ vs. $SEQ = 0.789$), $\chi^2(1) = 3.12, P = 0.077$. Thus, eliminating these 65 eyewitnesses reduced statistical power without having an appreciable effect on the pattern of results.

As indicated earlier, a five-member research team recoded the presence vs. absence of corroborating evidence based on its judgment of what counted as evidence. When the recoded corroborating evidence data from all of the witnesses were analyzed, the results continued to show a trend favoring the simultaneous procedure ($SIM = 0.912$ vs. $SEQ = 0.804$), $\chi^2(1) = 2.77, P = 0.096$. However, when the reduced recoded data set was analyzed (eliminating 65 witnesses based on their questionnaire responses), the effect, although continuing to favor the simultaneous procedure ($SIM = 0.920$ vs. $SEQ = 0.842$), was no longer marginally significant, $\chi^2(1) = 1.30, P = 0.244$. Although not significant, even for this analysis, more suspects identified from simultaneous lineups had independent corroborating evidence of guilt compared with sequential lineups ($SIM = 46$ vs. $SEQ = 32$), pointing to possible guilt, and fewer had no evidence of guilt ($SIM = 4$ vs. $SEQ = 6$), pointing to possible innocence. It therefore seems fair to conclude that all of these corroborating evidence analyses at least weigh against the notion that sequential lineups are diagnostically superior to simultaneous lineups. To the extent that these findings are interpreted as supporting the diagnostic superiority of the simultaneous procedure, they are consistent with the statistically significant corroborating evidence findings from the recent Austin police department field study (22).

Finally, we fit the equal variance signal detection model, with $\mu_{Target}$ fixed at 0.35 (free parameters = $\mu_{Target}$, $CI$, $c_2$, and $c_3$), separately to the simultaneous and sequential Houston field data broken down by confidence (Table S1). When the full data set was analyzed, $\mu_{Target}$ was significantly higher for the simultaneous procedure than for the sequential procedure ($SIM = 2.87$ vs. $SEQ = 2.03$), $\chi^2(1) = 5.01, P = 0.025$. When the reduced data set was analyzed (excluding the 65 witnesses discussed above), the difference in the estimated value of $\mu_{Target}$ still favored the simultaneous procedure ($SIM = 2.74$ vs. $SEQ = 2.12$), but the effect was no longer marginally significant, $\chi^2(1) = 2.06, P = 0.15$. A similar pattern of results held true across a variety of approaches to modeling the data (see SI Results, Signal Detection Model Fits). Thus, it seems fair to conclude that the signal detection analyses weigh against the notion that sequential lineups are diagnostically superior to simultaneous lineups. To the extent that these findings are
interpreted as supporting the simultaneous procedure, they are consistent with recent laboratory-based ROC analyses (18–21).

**Discussion**

Our results suggest that, contrary to a widely held view that confidence and accuracy are only weakly related but in agreement with recent experimentally controlled noncrime studies using a calibration approach (9–11), eyewitness confidence appears to be a reliable indicator of accuracy when an identification is made from a police lineup. The strong relationship between confidence and accuracy is indirectly suggested by trends in the raw data (Fig. 1B) and is directly implied by model-based estimates (Fig. 4). In addition, and again contrary to a widely held view, the present results reinforce both ROC analyses of laboratory-based data (18–21) and another police department field study analysis (22) suggesting that sequential lineups are not diagnostically superior to simultaneous lineups and that the reverse is more likely to be true (although, depending on how the data were analyzed here, the simultaneous advantage was not always significant).

Critically, our conclusions apply only to fair lineups initially administered to adults in double-blind fashion, not necessarily to unfair lineups, contemporary lineups, lineups administered to children, or to any ID associated with a subsequent memory test (including the one that occurs much later in a court of law). It is well known that memory is malleable such that by the time a witness testifies at trial or pretrial hearings, an initial low-confidence ID can be transformed into a high-confidence ID (29). In light of the recent recommendations made by a committee of the National Academy of Sciences on eyewitness identification—specifically, that lineups should be administered in double-blind fashion and that initial eyewitness confidence should be recorded (25)—it seems likely that the double-blind approach will be increasingly used by law enforcement agencies and that eyewitness confidence statements will be increasingly available. Under those conditions, our findings suggest that eyewitness confidence is a highly reliable indicator of accuracy and that simultaneous lineups are, if anything, diagnostically superior to sequential lineups.

**Methods**

A more detailed description of the experimental design/methods is provided in SI Methods.

**Participants.** The participants were 45 police investigators in the Robbery Division of the Houston Police Department and 717 eyewitnesses who were presented with photo lineups between January 22 and December 5, 2013. Inclusion criteria were that (i) the robberies involved strangers and (ii) the witnesses had not previously viewed a photo spread with the suspect.

**Informed Consent.** The study was approved by Protection of Human Subjects Committee in the Office of Research and Sponsored Programs at Sam Houston State University (Protocol 2012-08-202). All of the investigators who participated in the study signed an informed consent document, and witnesses were provided with a cover letter that explained risks and their rights. In addition, at the conclusion of the ID procedure, a survey was provided to each witness asking how the photos were shown to them (all at once or one at a time), whether the detective could see which photos they were viewing, whether they picked someone from the photos, etc. If they completed and returned the survey to the detective, then they were agreeing to participate.

**Procedure.** Witnesses were pseudorandomly assigned to one of four photo lineup conditions: blind sequential (n = 161), blind simultaneous (n = 187), blinded sequential (n = 175), and blinded simultaneous (n = 194). A lineup contained six photos (one suspect and five fillers). For the simultaneous procedure, this was the eyewitnesses viewing six photos at the same time. For the sequential procedure, the six photos were viewed one at a time. In the blind procedure, an investigator with no knowledge of the suspect’s identity administered the lineup. In the blinded procedure, the primary investigator conducted the viewing but was prevented from knowing which photo the witness was viewing. Eyewitnesses who made suspect IDs or filler IDs from these lineups were asked to supply a confidence rating using a three-point scale. For each case, an investigating officer filled out a questionnaire that addressed a variety of issues pertaining to the case (e.g., where was the lineup conducted, is there independent evidence of suspect guilt, what was the level of confidence expressed by the eyewitness, etc.).

**ACKNOWLEDGMENTS.** The following individuals in the Houston Police Department played an instrumental role in the experiment: Executive Assistant Chief Martha Montalvo, Assistant Chief George Buenik, Lieutenant Heather Morris, Captain Mark Holloway, Captain Lori Bender, Sergeant Steve Morrison, and the men and women of the Robbery Division. This work was supported in part by the National Science Foundation under Grant SES-1456571 (to J.T.W.).
Supporting Information

Wixted et al. 10.1073/pnas.1516814112

SI Methods

The experiment was conducted in the Houston Police Department (HPD) Robbery Division because the largest volume of photo spread and lineup procedures are conducted by robbery investigators.

Experimental Conditions. The experiment was designed to test the outcomes of four different methods of showing photos and individuals to eyewitnesses: (i) blind simultaneous, (ii) blind sequential, (iii) blinded sequential, and (iv) blinded simultaneous. Each of these procedures was consistent with HPD eyewitness identification procedures at the time of the study and thus did not introduce new procedures.

With the simultaneous presentation method, the eyewitness viewed all photos or live lineup members at the same time. In the sequential presentation method, photos and live lineup members were viewed one at a time. In the blind procedure, the primary investigator who knew the identity of the suspect prepared the photo spread, but an investigator with no knowledge of the suspect’s identity administered the viewing with the witness. In the blinded procedure, the primary investigator conducted the viewing, but a mechanism was used to prevent the investigator from (i) knowing the suspect’s position in the photo spread or lineup and (ii) knowing which photo or individual the suspect was viewing.

For the blinded procedures, the primary investigator selected the suspect’s photo and the five filler photos. A different investigator then randomly ordered the photos and placed them into a folder (or folders) for the investigating officer to administer. For the sequential procedure, one photo was placed into each of six folders. The stack of folders was then provided to the investigating officer, who then administered the photos one at a time to the witness. During the procedure, the investigator was positioned so it was not possible to view the photos while the witness viewed the photos. The blinded simultaneous procedure was similar except that the investigator who randomly ordered the photos placed the simultaneous photospread into a single folder before providing it to the investigating officer (who then administered the photos simultaneously to the witness).

In late 2012, all HPD criminal investigators were trained on the experimental procedures for showing photo spreads and lineups to eyewitnesses. These cases were assigned to receive one of the four experimental procedures for showing photo spreads and lineups to eyewitnesses. These cases were assigned to a treatment condition because they generate investigative activities, such as lineups. The characteristics of each determined whether a live lineup or a video lineup would be conducted (e.g., prosecutors may request a live viewing when a suspect is in custody). Only a small percentage of cases involved live or video lineups, and they were not included in any of our analyses, which focused solely on the more commonly used photo spread lineup procedure.

When a robbery report is routed to the Robbery Division, it is automatically assigned an HPD case number (referred to as an “incident number” by HPD personnel) by the computerized information management system. The case number determined which experimental procedure was to be used. Each HPD case number is a unique identifier that includes information about each determined whether a live lineup or a video lineup would be conducted (e.g., prosecutors may request a live viewing when a suspect is in custody). Only a small percentage of cases involved live or video lineups, and they were not included in any of our analyses, which focused solely on the more commonly used photo spread lineup procedure. The study randomly assigned investigations to a treatment condition, not to identification procedures. This means that when there were multiple eyewitnesses to a crime, all of the eyewitnesses were tested using the same procedure (to avoid confusing attorneys, judges, and jurors if the case proceeded to trial). In addition, when several crimes were judged to consist of a connected series of related crimes, the cases were also treated as a single investigation. Therefore, all eyewitnesses in these cases were tested using the same procedure. This restriction on our research protocol was necessary to avoid any detrimental effects of our study on the judicial process, but it also introduced the risk of the four conditions being unbalanced with respect to key variables. Therefore, we performed a variety of manipulation checks described below.

Our analysis was limited to 717 photo spreads that satisfied the following criteria: (i) the robberies involved strangers, (ii) the eyewitnesses had not previously viewed a photo spread with the suspect, and (iii) the photo spreads followed the experimental protocol that should have been used during the investigation. The number of cases in each condition that satisfied these criteria were: blinded simultaneous (n = 194), blinded sequential (n = 175), blind simultaneous (n = 187), and blind sequential (n = 161).

SI Results

The raw frequency counts and proportions of SIDs and FIDs (broken down by confidence) and no IDs for the blind and blinded simultaneous and sequential conditions are shown in Table S1. For the blind condition, overall response bias was similar (e.g., the proportion of no IDs was nearly identical for simultaneous and sequential lineups). For the blinded condition, overall response bias appears to be higher for the sequential condition (e.g., the proportion of no IDs was considerably higher for simultaneous lineups than for sequential lineups). However, as described in more detail next, this likely reflects the fact that the blinded sequential condition ended up with more guilty suspects (hence more suspect choosing) than the simultaneous procedure.
Manipulation Checks. 

Experiment investigator survey. For each case, an investigating officer filled out a questionnaire that addressed many issues pertaining to the case (e.g., “Where was the lineup conducted?” “Is there independent evidence of suspect guilt?” “What was the level of confidence expressed by the eyewitness?”). The full questionnaire can be found in appendix A of ref. 24 (Eyewitness experiment investigator survey). The first step in the analysis determined the degree to which the randomization procedure generated four treatment groups that were roughly equivalent in terms of key variables. Table S2 shows the results of these 11 comparisons. One result that stands out concerns corroborating evidence (second line). For this questionnaire item, the investigating officer answered “Was any corroborating evidence available in this case?,” with the options being “Yes” or “No.” For three of the four conditions, the values fell in a fairly narrow range of 0.62–0.70, but, for the blinded sequential condition, the value was conspicuously higher (0.91). The difference in the proportion of lineups with corroborating evidence was highly significant, \( \chi^2(3) = 45.90, P < 0.00001 \). This difference remains significant if the alpha level is adjusted to control for the 11 comparisons shown in Table S2 (using the Bonferroni correction, the alpha level is 0.05/11 = 0.0045). If the existence of corroborating evidence is assumed to be a proxy for likely guilt, then this result suggests that the blinded sequential condition ended up with a much higher proportion of guilty suspects than the other three conditions, which were similar to each other. Fortunately, as described later, the fact that the equal variance signal detection model includes a base rate parameter \( p_{\text{Target}} \) means that the two lineup procedures can still be differentially evaluated despite the apparent difference in base rate of guilty suspects for the blinded simultaneous and sequential lineups.

Two other effects are also significant even after applying the Bonferroni correction for the inflation of Type I error. One effect is the “Location” in which the lineup was administered. This effect is entirely understandable and reflects the fact that police officers tended to conduct blind lineups (which required another officer) in a police facility. The other effect is “Interpreter Used.” There is no obvious reason why this variable would differ across conditions. The three apparent trends (e.g., for Witness Under Influence, Witness Saw Photo, and Witness not Wearing His/Her Glasses) were not significant using the Bonferroni-corrected alpha level of 0.0045. Nevertheless, in Results, Simultaneous vs. Sequential Lineups, we also performed comparisons between the blind simultaneous and sequential lineups after excluding the 65 witnesses who fell into these categories. When that was done, the simultaneous lineup advantage was still apparent but was no longer significant.

Lineup fairness. To examine the potential differences between the four experimental groups, the fairness of photo spreads was examined for a sample of 60 randomly selected photo spreads. This analysis measures the degree to which the suspect “stands out” in the lineup by providing the lineups to mock witnesses to see if they could identify the suspect. Fifteen photo spreads were randomly selected from each of the four experimental conditions. The sample of mock witnesses included 15 HPD police cadets, 15 volunteers from the HPD Positive Interaction Program, and 19 Sam Houston State University students \((n = 49\) mock witnesses in all). Each mock witness was provided with the descriptions of the suspect and then attempted to identify that individual from the lineup. This measure of lineup fairness did not differ significantly for simultaneous lineups \((\text{mean} = 0.21)\) vs. sequential lineups \((\text{mean} = 0.25)\), but it did differ significantly for blinded lineups \((\text{mean} = 0.28)\) vs. blind lineups \((\text{mean} = 0.18)\). \( t(29) = 3.14, P < 0.01 \), but the mean of the blind lineups \((0.18)\) did not differ significantly from the expected value for a fair lineup, \( t(29) = 0.76 \). Three of the 30 blinded lineups had values that fell more than 2 SDs below the mean of 0.28 (biased away from the suspect), whereas 15 had values that fell more than 2 SDs above the mean of 0.28 (biased toward the suspect). By contrast, 4 of the 30 blind lineups had values that fell more than 2 SDs below the mean of 0.18 (biased away from the suspect), whereas 7 had values that fell more than 2 SDs above the mean of 0.18 (biased toward the suspect). The fact that the blinded lineups were, on average, unfair (biased toward the suspect) is another reason why our primary analysis focused on the blind lineups. The difference in lineup fairness between the blind and blinded conditions may indicate that police investigators who are aware that the lineup they have constructed will soon be scrutinized by a different investigator (the one who will be asked to administer it) take greater care to ensure that they prepare a fair lineup.

High-Threshold Model. The derivation of Eqs. 5 and 6 is straightforward. Adding Eqs. 2 and 4 yields

\[
F_n = n_{TP} \left[ (1-p) \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] + n_{TA} \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] \right].
\]

Setting \( n_{TP} = n_{TA} = n \), where \( n = N/2 \) (i.e., assuming equal base rates), we can algebraically solve for \( p \):

\[
F_n = n \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] + n \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] \]

\[
F_n = n \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] \cdot \left[ 1 - p \cdot n \cdot \left( \frac{5}{6} \right) \right] + n \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] \]

\[
F_n = 2 \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] - p \cdot n \cdot g \cdot \left( \frac{5}{6} \right) \]

\[
p \cdot n \cdot g = 2 \cdot \left[ g \cdot \left( \frac{5}{6} \right) \right] \]

\[
p = \frac{1}{n} \left[ g \cdot \left( \frac{5}{6} \right) \right] \]

\[
p = 2 \cdot \left( \frac{5}{6} \right) \cdot F \left( \frac{n}{g} \right)
\]

[S1]

Next, we can use Eq. S1 to solve for \( g \). Adding Eqs. 1 and 3 and setting \( n_{TP} = n_{TA} = n \) yields

\[
S_n = n \cdot \left[ (1 - p) \cdot g \cdot \left( \frac{1}{6} \right) \right] + n \cdot g \cdot \left( \frac{1}{6} \right)
\]

\[
S_n = n \cdot \left[ (1 - p) \cdot g \cdot \left( \frac{1}{6} \right) \right] + n \cdot g \cdot \left( \frac{1}{6} \right)
\]

Substituting the expression for \( p \) given by Eq. S1 into this expression yields a rather complex equation that is easily simplified,

\[
S_n = n \cdot \left[ 2 \cdot \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right]
\]

\[
+ \left[ 1 - 2 \cdot \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right] \cdot g \cdot \left( \frac{1}{6} \right)
\]

\[
+ n \cdot g \cdot \left( \frac{1}{6} \right)
\]

\[
S_n = n \cdot \left[ 2 \cdot \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right]
\]

\[
+ \left[ 1 + \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right] \cdot g \cdot \left( \frac{1}{6} \right) + n \cdot g \cdot \left( \frac{1}{6} \right)
\]

\[
S_n = n \cdot \left[ 2 \cdot \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right] + n \cdot g \cdot \left( \frac{1}{6} \right)
\]

\[
+ \frac{F}{5} + n \cdot g \cdot \left( \frac{1}{6} \right)
\]

\[
S_n = n \cdot \left[ 2 \cdot \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right] + F/5
\]

\[
S_n = n \cdot \left[ 2 \cdot \frac{n \cdot g}{\left( \frac{5}{6} \right)} \right] + F/5
\]

This expression can now be solved for \( g \),

\[
S_n \cdot 2F/5 = 6F/5\cdot g
\]

\[
S_n \cdot 2F/5 = 6F/5\cdot (1/g)
\]

\[
g = (6F/5)/(S_n \cdot 2F/5)
\]

Multiplying the numerator and denominator by \( -5 \) yields
This is Eq. 5. Substituting Eq. S2 for $g$ in Eq. S1 allows us to solve for $p$,

$$p = \frac{2 \cdot 6 \cdot F}{(5 \cdot n \cdot (6 \cdot F) / (10 \cdot n - 5 \cdot S + F))}$$

$$p = \frac{2 \cdot 6 \cdot F}{(5 \cdot n \cdot [10 \cdot n - 5 \cdot S + F]) / (6 \cdot F)}$$

$$p = 2 \cdot (10 \cdot n - 5 \cdot S + F) / (5 \cdot n)$$

$$p = (10 \cdot n \cdot 10 \cdot n + 5 \cdot S + F) / (5 \cdot n)$$

$$p = (5 \cdot S - F) / (5 \cdot n).$$

This is Eq. 6 in the main text.

Once $g$ and $p$ are known, they can be substituted into Eqs. 1 and 3 to estimate the number of suspect IDs from target-present and target-absent lineups ($n_{TP}$ and $n_{TA}$, respectively), and these values can be used to estimate suspect ID accuracy, which equals $n_{TP} / (n_{TP} + n_{TA})$.

The equations for $g$ and $p$ (Eqs. 5 and 6, respectively) allow one to compute suspect ID accuracy for all IDs, but one would also like to use the high-threshold model to compute suspect ID accuracy separately for each level of confidence. To do so, the computational steps involved in computing suspect ID accuracy are identical, except that, now, confidence-specific values for $S$ and $F$ are used to compute $g$ and $p$ from Eqs. 5 and 6 (the value of $n$ in those equations remains unchanged). However, making this move implies additional parameters that allow for different levels of confidence to be expressed when in the above threshold state (which occurs with probability $p$) or in the guessing state (which occurs with probability $g$). For example, the probability of a high-confidence guess is given by

$$g \cdot c_{High} = \frac{(6 \cdot F_{High})}{(10 \cdot n - 5 \cdot S_{High} + F_{High})}$$

[S3a]

where $c_{High}$ is the probability of expressing high confidence in an ID despite being in the guessing state. The probability of instead expressing medium or low confidence in an ID despite being in the guessing state are $c_{Med}$ and $c_{Low}$, such that

$$g \cdot c_{Med} = \frac{(6 \cdot F_{Med})}{(10 \cdot n - 5 \cdot S_{Med} + F_{Med})}$$

[S3b]

$$g \cdot c_{Low} = \frac{(6 \cdot F_{Low})}{(10 \cdot n - 5 \cdot S_{Low} + F_{Low})}$$

[S3c]

where $c_{High} + c_{Med} + c_{Low} = 1$.

Similar considerations apply to Eq. 6 such that

$$p \cdot d_{High} = \frac{(5 \cdot S_{High} + F_{High})}{(5 \cdot n)}$$

[S4a]

$$p \cdot d_{Med} = \frac{(5 \cdot S_{Med} + F_{Med})}{(5 \cdot n)}$$

[S4b]

$$p \cdot d_{Low} = \frac{(5 \cdot S_{Low} + F_{Low})}{(5 \cdot n)}$$

[S4c]

where $d_{High}$, $d_{Med}$, and $d_{Low}$ represent the probability of expressing high, medium, or low confidence in a guilty suspect ID despite being in the above-threshold (i.e., recognition) state, and $d_{High} + d_{Med} + d_{Low} = 1$.

Note that the values on the right side of each of these equations are known, so it is possible to directly compute the six combined parameter values on the left side of these equations. That is, one can directly compute these six parameters ($\tau_1$ through $\tau_6$).

There are only 6 degrees of freedom in the data, so this is as far as one can go. There are 6 degrees of freedom in the data because those data fall into seven cells: three confidence-specific suspect ID counts, three confidence-specific filler ID counts, and the number of no IDs. Still, these six combined parameter values are all that are needed to compute confidence-specific estimates of suspect ID accuracy. For example, to compute high-confidence suspect ID accuracy, $g \cdot c_{High}$ is first computed from Eq. S3a.

There were 17 high-confidence filler IDs ($F_{High} = 17$), 71 high-confidence suspect IDs ($S_{High} = 71$), and, assuming equal base rates, $348 / 2 = 174$ target-present lineups and 174 target-absent lineups ($n = 174$). Thus, according to Eq. S3a, $g \cdot c_{High} = 0.07$. According to Eq. S4a, $p \cdot d_{High} = 0.39$. To estimate the number of correct and incorrect suspect IDs made with high confidence, one simply substitutes $g \cdot c_{High}$ for $g$ in Eqs. 1 and 3 and substitutes $p \cdot d_{High}$ for $p$ in Eq. 1. With $n_{TP} = n_{TA} = n$, the estimated correct and incorrect high-confidence suspect IDs come to 68.9 and 2.1, respectively, so estimated suspect ID accuracy was 68.9 / (68.9 + 2.1) = 0.97.

Signal Detection Model Fits. The signal detection model illustrated in Fig. 2 uses a simple decision rule according to which an ID is made if the most familiar person in a lineup exceeds the lowest confidence criterion ($c1$), with confidence (low, medium, or high) being determined by the highest criterion that is exceeded. The signal detection model was fit to the experimentally controlled field data and the Houston Police Department field data via Monte Carlo simulation.

First, initial values were set for each of the model parameters ($\mu_{Target}$, $\sigma_{Target}$, $c1$, $c2$, and $c3$). The values of $\mu_{Lure}$ and $\sigma_{Lure}$ were always fixed at 0 and 1, respectively. Next, for each of $n_{TP}$ simulated witnesses (where $n_{TP}$ is the number of target-present lineups used in the experiment being fit), a predicted target-present decision was made by randomly drawing five values from the lure distribution, randomly drawing one value from the target distribution, and then applying the decision rule described above. Once all $n_{TP}$ trials were completed, one set of predicted target-present data had been generated. Similarly, for each of $n_{TA}$ simulated witnesses (where $n_{TA}$ is the number of target-absent lineups used in the experiment being fit), a predicted target-absent decision was made by randomly drawing six values from the lure distribution, and then applying the decision rule described above. Once all $n_{TA}$ trials were completed, one set of predicted target-absent data had been generated. These predicted target-present and target-absent data constituted results from one simulated experiment. However, because the simulation involves stochastic processes, the predicted data from a single simulated experiment are noisy. Thus, for each iteration of a fit, predicted values were obtained by running 500 simulated experiments and then averaging the results. Those averaged results were compared with the observed data (computing a $\chi^2$ goodness-of-fit statistic), at which point the parameters were adjusted and the next iteration of the fit occurred. Using fminsearch in MATLAB, the fitting process continued until the $\chi^2$ was minimized.
Fitting the model to the Palmer et al. (11) experimentally controlled field data. As noted in the main text, in a large-scale \((n = 908)\) investigation into the relationship between confidence and accuracy (11), experimenters approached participants in parks and shopping malls and asked them to view a target person (the “perpetrator”). Approximately half the participants were tested using a target-present lineup and the other half using a target-absent lineup. Thus, the base rate of target-present lineups was known to be ~50%. We first fit the five-parameter unequal variance signal detection model (parameters are \(\mu_{\text{Target}}, \sigma_{\text{Target}}, c_1, c_2,\) and \(c_3\)) to the uncollapsed data shown in Fig. 3A (except that suspect and filler IDs from target-present lineups and filler IDs from target-absent lineups were broken down by three levels of confidence; no estimated frequency counts were involved in the fit). This is how the model would ordinarily be fit to empirical data where it is known which trials involved target-present lineups and which involved target-absent lineups. For a given set of parameter values, the model generates predicted data separately for target-present and target-absent lineups. Thus, for each iteration of this fit, the predicted data were assessed by comparing the predicted values to the uncollapsed observed data in Fig. 3A. The results of the fit are shown in Table S3. The fit was very good, \(\chi^2(4) = 4.25\). Unlike what is typically found in basic list memory studies, an equal variance model turned out to be sufficient, i.e., \(\sigma_{\text{Target}}\) did not differ appreciably from 1, thus \(\sigma_{\text{Target}} \approx \sigma_{\text{Lame}}\).

We next fit the model to the collapsed experimentally controlled field data shown in Fig. 3C. For this fit, observed suspect IDs and filler IDs from target-present lineups were combined with estimated suspect IDs and estimated filler IDs from target-absent lineups. As noted in the main text, on each iteration of the fit, the predicted data were combined across target-present and target-absent lineups (thereby losing predicted information specific to lineup type) and the \(\chi^2\) goodness-of-fit statistic was computed by comparing the collapsed predicted data to the collapsed observed data from ref. 11. In other words, we fit the model to these data as if we were fitting it to police department field data, where target-present and target-absent lineups are combined. The model-fitting procedure assumed that target-present and target-absent lineups had been combined in equal proportion (i.e., as if the experimenter had equal base rates, which is known to be true of the experimentally controlled field study), and the fit was very good, \(\chi^2(1) = 0.34\) (see predicted values Fig. 3D and E). The results in Table S3 show that, once again, an equal variance model turned out to be sufficient. Indeed, and remarkably, the estimated parameter values were nearly identical whether the model was fit to the uncollapsed data or to the collapsed data. These results indicate that the model is capable of recovering information about target-present and target-absent lineups even when the data are collapsed (albeit in known proportions) across lineup type. Even more remarkably, as described next, if an equal variance model is assumed, the model can also recover the true underlying base rate when the target-present and target-absent lineups are combined in an unknown proportion.

Validating base rate signal detection analyses. We tested the ability of the signal detection model to recover the underlying base rate of target-present lineups using the experimentally controlled field data (11). For the base rate analyses described here, we fixed \(\sigma_{\text{Target}} = \sigma_{\text{Lame}} = 1\), and we added a target base rate parameter to the model (so its five parameters now were \(\mu_{\text{Target}}, \sigma_{\text{Target}}, c_1, c_2,\) and \(c_3\), where \(p_{\text{Target}}\) estimates \(n_T / (n_T + n_{\text{Lame}})\). When this model was fit to the collapsed data in Fig. 3C, the estimated value of \(p_{\text{Target}}\) was 0.51, very close to the true target-present base rate. Next, we mixed the target-present and target-absent data in varying proportions, fitting the signal detection model each time to the collapsed data (aggregated across target-present and target-absent lineups) to see if it could recover the true base rate value. Indeed, the estimated value of \(p_{\text{Target}}\) very closely tracked the true underlying base rate (Fig. S1). These data suggest that, if the equal variance model is assumed to also apply to the Houston field data, it can be used to accurately estimate the base rate of target-present lineups in that study (i.e., it can be used to estimate this real-world base rate). As noted in Results, A Model-Based Estimate of the Target-Present Base Rate and as described in more detail next, when the equal variance model was fit to the Houston field data, the value of \(p_{\text{Target}}\) was estimated to be 0.35. This was true for the separate fits of the model to the blind simultaneous and blind sequential data. In other words, despite being independent fits, they yielded the same estimate of \(p_{\text{Target}}\). It is important to stress that this base rate estimate depends on the validity of the equal variance model that was fit to the data. If future research suggests that an unequal variance model is more likely to characterize lineup data, then the base rate estimate would change accordingly.

Fitting the model to the blind Houston field data. We first fit a five-parameter unequal variance signal detection model (parameters are \(\mu_{\text{Target}}, \sigma_{\text{Target}}, c_1, c_2,\) and \(c_3\)) separately to the blind simultaneous and blind sequential data from the Houston field study. Because the true base rate of target-present lineups is unknown, we performed the fit three times, assuming a 25% base rate, 50% base rate, and 75% base rate. We also fit the model to the data combined across simultaneous and sequential lineups. The results of these fits are shown in Table S4. The \(\chi^2\) goodness-of-fit statistics show that, as a general rule, the model fit very well. Moreover, a discriminability estimate \((d')\) shows that, generally speaking, the simultaneous procedure outperformed the sequential procedure.

However, as noted in Results, Simultaneous vs. Sequential Lineups, 65 witnesses reported that they (i) encountered a photo of the suspect before being presented with the photo lineup, (ii) were under the influence of alcohol when they witnessed the crime, and/or (iii) were not wearing their prescribed glasses during the crime. The results of these model fits on the reduced data set (eliminating these 65 witnesses) are shown in Table S5. The fits are still good, but now the apparent \(d'\) advantage for simultaneous lineups is limited to the low base rate condition and even slightly reverses at the high (75%) base rate condition. Note that the predicted confidence-accuracy relationship was virtually identical whether these 65 witnesses were included in the analysis or not.

As described in the main text, based on the results of the model fit to the experimentally controlled field data (which suggested an equal variance model; see Table S3), we next made the assumption that an equal variance model \((\sigma_{\text{Target}} = \sigma_{\text{Lame}})\) also applies to the blind lineup data from the Houston field study, thereby allowing us to add a new parameter to the model. As noted above, the new parameter, \(p_{\text{Target}}\), provides an estimate of the base rate of target-present lineups. Model-based comparisons between the blind simultaneous and sequential lineups were performed by fitting the model to the simultaneous and sequential data concurrently, estimating the following five parameters for each lineup procedure: \(\mu_{\text{Target}}, c_1, c_2, c_3,\) and \(p_{\text{Target}}\). These five parameters were allowed to differ for the simultaneous and sequential lineups, so there were 10 free parameters in all. There were 12 degrees of freedom in the data (6 per lineup type), so this was a 2-degree-of-freedom fit. The fit was very good, \(\chi^2(2) = 1.35, P = 0.51\).

We next eliminated three parameters by constraining \(c_1, c_2,\) and \(c_3\) to be equal to the two lineup formats. The resulting change in the goodness-of-fit \(\chi^2\) was not significant, \(\chi^2(3) = 1.94, P = 0.59\), indicating that these parameters did not differ for the simultaneous and sequential formats. Thus, \(c_1, c_2,\) and \(c_3\) were constrained to be equal across lineup format for the subsequent fits. Next, we eliminated another parameter by constraining \(p_{\text{Target}}\) to be equal for the two lineup formats. Even when free to differ, the estimated value of this parameter was 0.35 for both lineup formats, so the goodness-of-fit \(\chi^2\) did not change at all when this constraint was added. Thus (obviously), the estimated value of \(p_{\text{Target}}\) did not differ for the
simultaneous and sequential formats. Finally, we tried eliminating one additional parameter by constraining \( \mu_{target} \) to be equal for the two lineup formats. This constraint resulted in a far worse fit, \( \chi^2(1) = 8.12, P = 0.004 \), indicating that the mean of the target distribution differed significantly for the simultaneous and sequential lineup formats. With all other parameters constrained to be equal, \( \mu_{target} \) was estimated to be 2.00 for the blind simultaneous lineups and 2.03 for the blind sequential lineups. In other words, simultaneous lineups were estimated to be diagnostically superior to sequential lineups (i.e., the ability to discriminate innocent from guilty suspects was higher for the simultaneous procedure). The estimated values of \( \mu_{Target} \) fit reported in Results. Simultaneous vs. Sequential Lineups (SIM = 2.87 vs. SEO = 2.03) differed slightly only because, for that fit, \( c1, c2, \) and \( c3 \) were allowed to differ for the two lineup formats. The results are very similar either way. However, as noted in the main text, 65 witnesses reported that they (i) encountered a photo of the suspect before being presented with the photo lineup, (ii) were under the influence of alcohol when they witnessed the crime, and/or (iii) were not wearing their prescribed glasses during the crime. When these witnesses are eliminated from the model fits, the difference in \( \mu_{Target} \) still favored the simultaneous procedure (SIM = 2.59 vs. SEO = 2.20), but the difference was no longer significant (\( P = 0.26 \)).

**Lap 1 vs. Lap 2 Choosing in the Blind Sequential Condition.** When the sequential lineup is used in mock crime laboratory studies, participants are typically permitted to view the photos only once (usually with the first ID being the one that counts). In actual practice, the police permit second viewings if the witness requests it. This was also true of the Houston Police Department field study. Investigators were instructed to allow witnesses to view sequential photo spreads a second time only if the witness requested to view the photos again. In other words, investigators did not give this option up front when giving the instructions. If the witness requested to view the photo spread or they requested to view a single photo again, the investigator showed the entire set of photos a second time, in the same order as the first showing. Thus, we can examine the data for witnesses who viewed the photos only once (1 lap) to see how the results compared with witnesses who viewed the photos more than once (2+ laps).

Investigators were asked to complete a survey item that asked how many times the witnesses viewed the photo spread. For 24 of the 161 blind sequential photo spreads, this information was not recorded, and we assumed that this was because only one viewing occurred. This counting rule resulted in 96 witnesses viewing the photos once and 65 viewing the photos more than once (59 twice, 5 three times, and 1 four times). In other words, 40% of the witnesses (65 out of 161) requested more than one viewing. For the analyses discussed next, we consider the data separately for those who viewed the photo spread once vs. those who viewed it more than once. The relevant frequency counts and proportions are presented in Table S6.

Responding was considerably more conservative for those who viewed the sequential photo spread only once compared with those who viewed it more than once. For example, for the lap 1 witnesses, 45 out of 96 made no ID (0.47), whereas for the lap 2+ witnesses, only 14 out of 65 made no ID (0.22), a difference that was highly significant, \( \chi^2(1) = 10.72, P = 0.0011 \). For those who made an ID, lap 1 witnesses identified a suspect 26 times and identified a filler 25 times (0.51 suspect IDs), whereas lap 2+ witnesses identified a suspect 20 times and identified a filler 31 times (0.39 suspect IDs). On the surface, it appears that the lap 2+ witnesses were less accurate, but the fact that they were also more liberal in their responding means that the two numbers cannot be meaningfully compared to assess relative discriminability. Instead, a measure that is not influenced by response bias (e.g., \( d' \)) is needed.

We fit the equal variance signal detection model, with \( \rho_{Target} \) fixed at 0.35 (free parameters are \( \mu_{Target}, c1, c2, \) and \( c3 \)), separately to the lap 1 and lap 2+ sequential Houston field data broken down by confidence (i.e., we fit the model to the frequency data shown in Table S6). With all parameters free to vary, the model fit the data well, \( \chi^2(4) = 1.50, P = 0.827 \). The estimated value of \( \mu_{Target} \) (i.e., \( d' \)) was actually higher, not lower, for the lap 2+ sequential witnesses (\( \mu_{Target} = 2.24 \)) compared with the lap 1 sequential witnesses (\( \mu_{Target} = 1.92 \)). However, the difference was not remotely close to being significant, \( \chi^2(1) = 0.35, P = 0.56 \). In other words, the large apparent difference in \( \mu_{Target} \) is an illusion because the fit is scarcely affected by constraining the parameters to be equal, in which case the estimated value of \( \mu_{Target} \) was 2.02. Thus, with these data, no difference in discriminability was detected between the lap 1 and lap 2+ witnesses. However, the other parameter estimates suggest that the lap 2+ witnesses may have been much more liberal about making an ID in general (i.e., \( c1 \) was noticeably lower for the lap 2+ witnesses) while being more conservative about making a high-confidence ID (i.e., \( c3 \) was noticeably higher for the lap 2+ witnesses). The estimates for \( c1, c2, \) and \( c3 \) were, respectively, 1.43, 1.78, and 2.18 for the lap 1 witnesses and 0.90, 1.62, and 2.52 for the lap 2+ witnesses. Indeed, when the \( c1, c2, \) and \( c3 \) parameters were constrained to be equal across the lap 1 and lap 2+ sequential witnesses, the fit was significantly worse, \( \chi^2(3) = 19.64, P < 0.001 \). This result indicates that the main difference between the lap 1 and lap 2+ witnesses is in their placement of the various decision criteria in the threshold, not in the ability to discriminate innocent from guilty suspects. It also indicates that the trends favoring the simultaneous procedure discussed earlier did not arise merely because the lap 2+ witnesses were included in the sequential analysis.

**Analyses of Blinded Lineups.** The data from the blinded simultaneous and sequential conditions were analyzed in the same way that the data from the blind conditions were analyzed in Results. Here, we first describe the basic empirical trends in the data and then present signal detection analyses of the data. The results of these analyses are all similar to the results from the analyses of the blind lineups, but various complexities involved with this data set (unbalanced corroborating evidence for simultaneous and sequential formats before any IDs being made, unfair lineups, and poor signal detection model fits) warrant caution in interpreting the results. We present the results for the sake of completeness and to underscore the fact that the results are similar to the results presented above and therefore offer no reason to question the interpretation of the blind lineup data.

**Confidence in suspect IDs and filler IDs from blind lineups.** For the blinded lineups (collapsed across simultaneous and sequential), the results were similar to the results observed for the blind lineups shown in Fig. 1A–D. Suspect IDs and filler IDs occurred with approximately equal frequency, whereas “no IDs” occurred with somewhat greater frequency (Fig. S2A). As with the blind data, most filler IDs were made with low confidence and most suspect IDs were made with high confidence (Fig. S2B). In other words, the proportion of IDs for each level of confidence that were suspect IDs—that is, suspect IDs (suspect IDs + filler IDs)—increased dramatically with confidence (Fig. S2C). For suspect IDs, the proportion of cases with corroborating evidence of guilt also increased as confidence in the ID increased (Fig. S2D).

**Fitting the model to the blinded Houston lineup data.** As with the blind data, we first fit a five-parameter unequal variance signal detection model (parameters = \( \mu_{Target}, \rho_{Target}, c1, c2, \) and \( c3 \)) separately to the blinded simultaneous and blinded sequential data from the Houston field study. Because the true base rate of target-present lineups is unknown, we performed the fits three times, assuming a 25% base rate, 50% base rate, or 75% base rate. We also fit the model to the data combined across simultaneous and sequential lineups. The results of these fits are shown in Table S7. The \( \chi^2 \) goodness-of-fit statistics show that, as
a general rule, the model fit the simultaneous data very well but provided a poor fit to the sequential data. It is not clear why the data from the blinded sequential condition are so poorly fit by the model, but the parameter estimates for the fits described below generally seem sensible and interpretable despite the poor fit to that condition.

**Base rate estimates.** Based on the results of the model fit to the experimentally controlled field data (which suggested an equal variance model), and again following the same steps we followed for the blind lineup data, we next made the assumption that an equal variance model ($\sigma_{\text{Target}} = \sigma_{\text{Lure}}$) also applies to the blinded Houston field data. Making this assumption allowed us to obtain an estimate of the base rate of target-present lineups ($P_{\text{Target}}$) separately for the simultaneous and sequential lineups. Model-based comparisons between the blinded simultaneous and sequential lineups were performed by fitting the data from both lineup procedures concurrently.

The first model included the following parameters: $\mu_{\text{Target}}$, $c_1$, $c_2$, $c_3$, and $P_{\text{Target}}$ which were all allowed to differ for simultaneous and sequential lineups (thus, there were 10 parameters in all). Mainly because the equal variance model was not able to accurately characterize the blinded sequential data (as discussed above), the simultaneous fit of the full model to the blinded simultaneous and sequential data were poor, $\chi^2(12) = 12.09$, $P = 0.049$. We next investigated the effect of constraining the parameters to be equal across lineup formats. Unlike the model fits to the blind simultaneous and sequential data, when the confidence criteria $c_1$, $c_2$, and $c_3$ were constrained to be equal across the blinded simultaneous and sequential lineup formats, the fit was worse, $\chi^2(3) = 7.85$, $P = 0.049$. The difference was barely significant, and given the number of significance tests performed, it might be regarded as a Type I error. Thus, we conservatively constrained the confidence parameters to be equal despite this barely significant result.

Next, $P_{\text{Target}}$ was constrained to be equal for the two lineup formats. Unlike the model fits to the blind simultaneous and sequential data, the fit was significantly worse, $\chi^2(6) = 32.94$, $P < 0.0001$. In other words, $P_{\text{Target}}$ was significantly greater for blinded sequential lineups compared with blinded simultaneous lineups (0.23 for simultaneous and 0.43 for sequential). Remarkably, the fact the estimated value of $P_{\text{Target}}$ was significantly greater for blinded sequential lineups is consistent with the entirely independent corroborating evidence estimates discussed earlier, which also suggest that more guilty suspects ended up in blinded sequential lineups than in blinded simultaneous lineups. In other words, two entirely independent analyses converge on the notion that an unusually high number of guilty suspects ended up in the blinded sequential condition (for reasons that are unclear).

**Model-based confidence—accuracy analyses.** The best-fitting unequal variance model to the combined data (Table S7) yielded the confidence—accuracy predictions shown in Fig. S2E (averaged across simultaneous and sequential lineups), which closely resemble the corresponding predictions for the blind data shown in Fig. 4A. Similarly, the confidence—accuracy relationship predicted by this best-fitting equal variance model (again averaged across simultaneous and sequential lineups) with $P_{\text{Target}}$ fixed at the estimated values described above exhibits a strong relationship between the confidence associated with a suspect ID and the accuracy of that ID (Fig. S2F), just as was true of the lineups from the blind condition (Fig. 4B). Thus, with regard to the relationship between confidence and accuracy, the results are very similar for the blind and blinded lineups.

**Discriminability for blinded simultaneous and sequential lineups.** For the next fit, we compared an equal variance model with $c_1$, $c_2$, $c_3$, and $P_{\text{Target}}$ constrained to be equal across lineup format (with $P_{\text{Target}}$ free to vary) to an equal variance model with $c_1$, $c_2$, and $c_3$ constrained to be equal across lineup format (with $P_{\text{Target}}$ and $P$ free to vary). This test asks whether discriminability differs for the blinded simultaneous and sequential lineup formats and is therefore a key test. The result indicated a significant simultaneous advantage, $\chi^2(1) = 4.31, P = 0.038$. The estimates of $P_{\text{Target}}$ for the blinded simultaneous and sequential lineups were 3.10 and 2.27, respectively. Thus, as with the fits to the blind lineups, the fits to the blinded lineup data suggest a discriminability advantage for simultaneous lineups.

Based on the values shown in Table S7, the argument could be made that the $c_1$ confidence parameter should be free to differ when comparing $P_{\text{Target}}$ for the blinded simultaneous and sequential lineups (whereas they were constrained to be equal across lineup format in the test described above). When that is done, the estimate for $P_{\text{Target}}$ still favors the simultaneous procedure (SIM = 2.90, SEQ = 2.48), but the difference is now only marginally significant, $\chi^2(1) = 2.88, P = 0.090$.

**Corroborating evidence of guilt analyses for simultaneous vs. sequential lineups.** Next, we describe a non-model-based comparison of simultaneous and sequential lineups based on corroborating evidence of guilt. Directly evaluating lineup performance based on evidence of guilt associated with identified suspects from blinded simultaneous and sequential lineups, as we did for the blind conditions, would not be useful here because it would obviously be higher for sequential lineups. However, some evidence of the relative diagnostic performance of the two lineup types can be obtained by examining whether corroborating evidence of guilt is higher for suspect IDs compared with no IDs within each lineup type. To the extent that suspect IDs for a given lineup type are associated with more corroborating evidence of suspect guilt compared with no IDs, eyewitness decisions would be diagnostic of guilt. In other words, this difference score (i.e., corroborating evidence against identified suspects minus corroborating evidence against suspects in lineups where no ID was made) provides an independent estimate of the accuracy to discriminate innocent from guilty suspects (an estimate that was directly quantified in the equal variance signal detection model fits summarized above).

For blinded simultaneous lineups, the proportion of cases with corroborating evidence against the suspect given that no ID was made was 0.54, whereas the proportion of cases with corroborating evidence against the suspect given that a suspect ID was made was 0.75, an increase that was significant, $\chi^2(1) = 5.00, P = 0.025$. Thus, by this measure, suspect IDs from the blinded simultaneous lineups were diagnostic of guilt. By contrast, for blinded sequential lineups, the proportion of cases with corroborating evidence against the suspect given that no ID was made was 0.93, whereas the proportion of cases with corroborating evidence against the suspect given that a suspect ID was made was 0.90. Thus, there was no indication of diagnosticity associated with suspect IDs in blinded sequential lineups. However, the high level of corroborating evidence against suspects across all blinded sequential lineups (0.91) would have made the detection of an increase associated with suspect IDs somewhat difficult. Thus, although these data are consistent with a simultaneous superiority effect (as was true of the equal variance signal detection fits described above), the findings are less compelling than those from the blind lineup analyses. On balance, it seems fair to conclude that for both the blind and the blinded lineups, (i) there is no evidence whatsoever for a sequential superiority effect, (ii) there are many indications of a simultaneous superiority effect, and (iii) the significance of the simultaneous superiority effect depends on the details of how the data are analyzed. In other words, there is no evidence of a sequential superiority effect, and, if anything, the data point to a simultaneous advantage (as would be predicted by recent laboratory-based ROC analyses).

**Recode Corroborating Evidence.** For the blind and blinded simultaneous and sequential lineups, five independent raters coded information in the fill-in text boxes where the investigating officer...
listed what the corroborating evidence was whenever it was
deemed to be present. For this work, the coders included one of
the authors (William Wells), two faculty members who have
significant experience conducting research with criminal inves-
tigators, a Ph.D. student who has research experience working
with criminal investigators, and a former Houston Police De-
partment robbery investigator (who has his Ph.D. and has done
work with interrater reliability checks). During this process, the
raters identified situations in which the investigator indicated
there was corroborating evidence but the team’s coding of the
text box information suggested it should not be counted as cor-
roborating evidence. The raters also identified situations in
which investigators included information in the fill-in text box
but nevertheless did not check the box to indicate there was
corroborating evidence.

For the independent coding, we used the coding that was rec-
ommended by three or more of the five coders. That is, if three raters
said the text box information should be counted as corroborating
evidence, then we counted it as such. The relationship between
corroborating evidence and confidence in a suspect ID is shown
in Fig. S2G. The results are similar to the results observed for
the blind lineups shown in Fig. 1D. The results for the recoded
blinded lineups are shown in Fig. S2H. Again, the relationship
remains essentially the same as for the originally coded data.

Fig. S1. Estimated value of $p_{\text{target}}$ (y axis) as a function of the true base rate of target-present lineups (x axis). The estimated values were based on a five-
parameter fit of the equal variance signal detection model to data from the experimentally controlled field study (11) that had been aggregated across target-
present and target-absent lineups in different proportions across fits. For example, at the lower left, the data were aggregated in such a way that 5% of the
data were from target-present lineups and 95% from target-absent lineups. Although some inaccuracy is apparent at very high base rates, for the most part,
the model is able to accurately recover the underlying base rate from data that have been aggregated across target-present and target-absent lineups. Thus,
all the model sees are the number of suspect IDs and filler IDs (across three levels of confidence) and no IDs. It has no information about how many of each
came from target-present lineups or target-absent lineups. Even so, it can recover the true base rate with remarkable accuracy, raising the possibility that it
might be able to do the same with police department field data (which are also aggregated across target-present and target-absent lineups in unknown
proportion).
Fig. S2.  (A) Frequency counts of eyewitness decision outcomes in the Houston field study for 194 blinded simultaneous and 175 blinded sequential lineups combined. (B) Frequency of SIDs and FIDs in A exhibited opposite trends as a function of confidence (low, medium, or high), χ²(2) = 69.2, P < 0.0001. (C) For ID made to a suspect or filler, the probability that it was a suspect IDs increased dramatically with confidence. (D) Proportion of suspect IDs rated by the investigating officer as having independent corroborating evidence of guilt increased with confidence in the ID. According to a Cochran–Armitage trend test, the effect was significant (one-tailed), Z = 2.80, P < 0.01. (E) Model-based estimates of the posterior probability of guilt associated with suspects identified from blinded lineups in the Houston field study for three different assumptions regarding base rates (BR) averaged across simultaneous and sequential lineups. (F) Model-based estimate of the posterior probability of guilt associated with suspects identified from blinded lineups assuming an equal variance signal detection model and including target-present base rate as a free parameter (estimated to be 0.23 for simultaneous lineups and 0.43 for sequential lineups). (G) For the blind lineups (recoded data), proportion of suspect IDs rated by the investigating officer as having independent corroborating evidence of guilt increased with confidence in the ID. According to a Cochran–Armitage trend test, the effect was significant (one-tailed), Z = 2.71, P < 0.01. (H) For the blinded lineups (recoded data), proportion of suspect IDs rated by the investigating officer as having independent corroborating evidence of guilt increased with confidence in the ID. According to a Cochran–Armitage trend test, the effect was marginally significant (one-tailed), Z = 1.42, P = 0.077. In all cases, error bars are SEs.

Table S1. Frequency counts and proportions of eyewitness decisions in the blind and blinded simultaneous and sequential conditions

<table>
<thead>
<tr>
<th>Type of ID</th>
<th>Frequency counts</th>
<th>Proportions</th>
<th>No ID proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Med</td>
<td>Low</td>
</tr>
<tr>
<td>Blind simultaneous</td>
<td>no ID</td>
<td>71</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>suspect ID</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Blind sequential</td>
<td>no ID</td>
<td>59</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>suspect ID</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>Blinded simultaneous</td>
<td>no ID</td>
<td>96</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>suspect ID</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>Blinded sequential</td>
<td>no ID</td>
<td>54</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>suspect ID</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>
Table S2. Differences and similarities between treatment groups

<table>
<thead>
<tr>
<th>Case characteristics</th>
<th>Blinded simultaneous n = 194</th>
<th>Blinded sequential n = 175</th>
<th>Blind simultaneous n = 187</th>
<th>Blind sequential n = 161</th>
<th>Statistical test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpreter used</td>
<td>6 (3%)</td>
<td>13 (7%)</td>
<td>16 (9%)</td>
<td>23 (14%)</td>
<td>$\chi^2 = 14.99 , P = 0.002$</td>
</tr>
<tr>
<td>Corroborating evidence</td>
<td>121 (62%)</td>
<td>160 (91%)</td>
<td>130 (70%)</td>
<td>105 (65%)</td>
<td>$\chi^2 = 45.90 , P = 0.000$</td>
</tr>
<tr>
<td>Witness under influence</td>
<td>17 (9%)</td>
<td>13 (7%)</td>
<td>3 (2%)</td>
<td>13 (8%)</td>
<td>$\chi^2 = 10.03 , P = 0.018$</td>
</tr>
<tr>
<td>Witness saw photo</td>
<td>21 (11%)</td>
<td>16 (9%)</td>
<td>23 (12%)</td>
<td>5 (3%)</td>
<td>$\chi^2 = 10.04 , P = 0.018$</td>
</tr>
<tr>
<td>Witness not wearing</td>
<td>7 (4%)</td>
<td>10 (6%)</td>
<td>6 (3%)</td>
<td>17 (11%)</td>
<td>$\chi^2 = 11.01 , P = 0.012$</td>
</tr>
<tr>
<td>his/her glasses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\chi^2 = 52.94 , P = 0.000$</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\chi^2 = 45.90 , P = 0.000$</td>
</tr>
<tr>
<td>Police</td>
<td>46 (24%)</td>
<td>34 (20%)</td>
<td>86 (46%)</td>
<td>51 (32%)</td>
<td>$\chi^2 = 6.98 , P = 0.007$</td>
</tr>
<tr>
<td>Facility</td>
<td>50 (26%)</td>
<td>61 (35%)</td>
<td>56 (30%)</td>
<td>47 (29%)</td>
<td>$\chi^2 = 1.83 , P = 0.001$</td>
</tr>
<tr>
<td>Home</td>
<td>48 (25%)</td>
<td>46 (26%)</td>
<td>30 (16%)</td>
<td>31 (19%)</td>
<td>$\chi^2 = 0.83 , P = 0.001$</td>
</tr>
<tr>
<td>Witness</td>
<td>40 (21%)</td>
<td>31 (18%)</td>
<td>11 (6%)</td>
<td>26 (16%)</td>
<td>$\chi^2 = 3.85 , P = 0.001$</td>
</tr>
<tr>
<td>Public setting</td>
<td>10 (5%)</td>
<td>3 (2%)</td>
<td>4 (2%)</td>
<td>6 (4%)</td>
<td>$\chi^2 = 0.83 , P = 0.001$</td>
</tr>
<tr>
<td>Serial investigation</td>
<td>82 (42%)</td>
<td>77 (44%)</td>
<td>76 (41%)</td>
<td>50 (31%)</td>
<td>$\chi^2 = 1.83 , P = 0.001$</td>
</tr>
<tr>
<td>Victim</td>
<td>166 (86%)</td>
<td>157 (90%)</td>
<td>166 (89%)</td>
<td>144 (89%)</td>
<td>$\chi^2 = 3.85 , P = 0.001$</td>
</tr>
<tr>
<td>Suspect in position 1</td>
<td>28 (14%)</td>
<td>19 (11%)</td>
<td>18 (10%)</td>
<td>13 (8%)</td>
<td>$\chi^2 = 0.83 , P = 0.001$</td>
</tr>
<tr>
<td>Good viewing opportunity</td>
<td>65 (34%)</td>
<td>68 (39%)</td>
<td>62 (33%)</td>
<td>55 (34%)</td>
<td>$\chi^2 = 0.83 , P = 0.001$</td>
</tr>
<tr>
<td>Weapon used</td>
<td>136 (70%)</td>
<td>118 (67%)</td>
<td>123 (66%)</td>
<td>109 (68%)</td>
<td>$\chi^2 = 0.83 , P = 0.001$</td>
</tr>
</tbody>
</table>

Because 11 post hoc statistical tests were performed, the Bonferroni-adjusted alpha level is 0.05/11 = 0.0045. Significant differences were observed for interpreter used, corroborating evidence, and location in which the lineup was administered.

Table S3. Parameter values of the signal detection model that minimize the $\chi^2$ goodness-of-fit statistic when the model is fit to the uncollapsed data shown in Fig. 3A and to the collapsed data shown in Fig. 3C

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Palmer et al. (11)</th>
<th>Palmer et al. (11) collapsed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c3$</td>
<td>2.57</td>
<td>2.58</td>
</tr>
<tr>
<td>$c2$</td>
<td>2.04</td>
<td>2.04</td>
</tr>
<tr>
<td>$c1$</td>
<td>1.51</td>
<td>1.51</td>
</tr>
<tr>
<td>$\mu_{\text{Target}}$</td>
<td>1.81</td>
<td>1.83</td>
</tr>
<tr>
<td>$\sigma_{\text{Target}}$</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>4.25</td>
<td>1.34</td>
</tr>
<tr>
<td>df</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>$p$</td>
<td>0.374</td>
<td>0.560</td>
</tr>
</tbody>
</table>

The uncollapsed data have 9 degrees of freedom, and the collapsed data have 6 degrees of freedom. For both model fits, five parameters were estimated. The model fit to the collapsed data assumed that target-present and target-absent lineups were combined in equal proportion (i.e., the base rate parameter, $p_{\text{Target}}$, was fixed to be 0.50). The similarity of the parameter estimates is striking.
### Table S4. Parameter values for the fit of the unequal variance model to the full blind data set

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Simultaneous</th>
<th>Sequential</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>c3</td>
<td>2.26</td>
<td>2.26</td>
<td>2.26</td>
</tr>
<tr>
<td>c2</td>
<td>1.74</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>c1</td>
<td>1.31</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>$\mu_{\text{Target}}$</td>
<td>2.54</td>
<td>2.33</td>
<td>1.30</td>
</tr>
<tr>
<td>$\sigma_{\text{Target}}$</td>
<td>0.28</td>
<td>1.85</td>
<td>2.45</td>
</tr>
<tr>
<td>$\chi^2(1)$</td>
<td>4.13</td>
<td>1.40</td>
<td>1.26</td>
</tr>
<tr>
<td>$p$</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td>$d_a$</td>
<td>3.46</td>
<td>1.57</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Parameter values of the unequal variance signal detection model that minimize the $\chi^2$ goodness-of-fit statistic when the model is fit to the full blind Houston field data set assuming target-present based rates of 25%, 50%, and 75%. Note that in order for the model to provide an acceptable fit, the estimated mean of the target distribution ($\mu_{\text{Target}}$) decreased and the estimated SD of the target distribution ($\sigma_{\text{Target}}$) increased as the base rate of target-present lineups increased. The suspect ID accuracy scores shown in Fig. 4A were derived from the best-fitting parameter values for the combined fits (simultaneous + sequential). For all fits, the mean and SD of the filler distribution were set to 0 and 1, respectively. The bottom row shows a discriminability estimate, $d_a = \mu_{\text{Target}}/\sqrt{(1 + \sigma_{\text{Target}}^2)/2}$; $d_a$ takes into account the unequal variance of the underlying distributions and is equal to $d'$ when the two variances are equal (i.e., when $\sigma_{\text{Target}} = 1$). For each fit, the data have 6 degrees of freedom, and the model has five free parameters (leaving 1 degree of freedom).

### Table S5. Parameter values for the fit of the unequal variance model to the reduced blind data set

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Simultaneous</th>
<th>Sequential</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>c3</td>
<td>2.44</td>
<td>2.44</td>
<td>2.45</td>
</tr>
<tr>
<td>c2</td>
<td>1.84</td>
<td>1.81</td>
<td>1.81</td>
</tr>
<tr>
<td>c1</td>
<td>1.36</td>
<td>1.38</td>
<td>1.38</td>
</tr>
<tr>
<td>$\mu_{\text{Target}}$</td>
<td>2.63</td>
<td>2.01</td>
<td>0.95</td>
</tr>
<tr>
<td>$\sigma_{\text{Target}}$</td>
<td>0.25</td>
<td>2.10</td>
<td>2.59</td>
</tr>
<tr>
<td>$\chi^2(1)$</td>
<td>0.47</td>
<td>1.47</td>
<td>1.26</td>
</tr>
<tr>
<td>$p$</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td>$d_a$</td>
<td>3.60</td>
<td>1.22</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Parameter values of the unequal variance signal detection model that minimize the $\chi^2$ goodness-of-fit statistic when the model is fit to the reduced blind Houston field data set (eliminating 65 witnesses based on their questionnaire responses about encountered a photo of the suspect, being under the influence of alcohol, and/or not wearing their prescribed glasses) assuming target-present based rates of 25%, 50%, and 75%. For all fits, the mean and SD of the filler distribution were set to 0 and 1, respectively. The bottom row again shows a discriminability estimate, $d_a = \mu_{\text{Target}}/\sqrt{(1 + \sigma_{\text{Target}}^2)/2}$. Note that the one case where the sequential procedure yields a slight advantage is in the 75% target-present base rate condition. For each fit, the data have 6 degrees of freedom, and the model has five free parameters (leaving 1 degree of freedom).
Table S6. Frequency counts and proportions of eyewitness decisions in the blind sequential condition, separated according to witnesses who viewed the photo spread only once and those who viewed it more than once

<table>
<thead>
<tr>
<th>ID type</th>
<th>Frequency counts</th>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Med</td>
</tr>
<tr>
<td>Lap 1 sequential no ID</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>suspect ID</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>filler ID</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Lap 2+ sequential no ID</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>suspect ID</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>filler ID</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

Table S7. Parameter values for the fit of the unequal variance model to the blinded data set

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Simultaneous</th>
<th>Sequential</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
</tr>
<tr>
<td>c3</td>
<td>2.40</td>
<td>2.40</td>
<td>2.40</td>
</tr>
<tr>
<td>c2</td>
<td>1.85</td>
<td>1.84</td>
<td>1.84</td>
</tr>
<tr>
<td>c1</td>
<td>1.44</td>
<td>1.44</td>
<td>1.43</td>
</tr>
<tr>
<td>(\mu_{\text{Target}})</td>
<td>2.71</td>
<td>1.21</td>
<td>0.33</td>
</tr>
<tr>
<td>(\sigma_{\text{Target}})</td>
<td>1.15</td>
<td>2.32</td>
<td>2.52</td>
</tr>
<tr>
<td>(\chi^2(1))</td>
<td>0.24</td>
<td>0.35</td>
<td>0.22</td>
</tr>
<tr>
<td>(p)</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Parameter values of the unequal variance signal detection model that minimize the \(\chi^2\) goodness-of-fit statistic when the model is fit to the blinded Houston field data assuming target-present based rates of 25%, 50%, and 75%. Note that in order for the model to provide an acceptable fit, the estimated mean of the target distribution (\(\mu_{\text{Target}}\)) decreased and the estimated SD of the target distribution (\(\sigma_{\text{Target}}\)) increased as the base rate of target-present lineups increased (trends that are also evident in the fits to the blind data in Table S4). The suspect ID accuracy scores shown in Fig. S2 were derived from the best-fitting parameter values for the combined fits (simultaneous + sequential). For all fits, the mean and SD of the filler distribution were set to 0 and 1, respectively. For each fit, the data have 6 degrees of freedom, and the model has five free parameters (leaving 1 degree of freedom).