The Reliability of Eyewitness Identifications from Police Lineups

John T. Wixted[a,1], Laura Mickes[b], John C. Dunn[c], Steven E. Clark[d], & William Wells[e]

[a]Department of Psychology, University of California, San Diego, La Jolla, California 92093, USA; [b]Department of Psychology, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, England; [c]School of Psychology, the University of Adelaide, North Terrace, Adelaide, 5005, Australia; [d]Department of Psychology, University of California, Riverside, Riverside, California, 92521, USA; [e]Department of Criminal Justice and Criminology, Sam Houston State University, Texas, 77341, USA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Laboratory-based mock-crime studies have often been interpreted to mean that (a) eyewitness confidence in an identification made from a lineup is a weak indicator of accuracy and (b) 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 U.S. 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 one-year period. Identifications were made using a 3-point confidence scale, and a signal-detection model was used to analyze and interpret the results. Our findings suggest that (a) confidence in an eyewitness identification from a fair lineup is a highly reliable indicator of accuracy and (b) if there is any difference in diagnostic accuracy between the two lineup formats, it likely favors the simultaneous procedure.

Eyewitness Identification | Confidence and Accuracy | vs. Sequential Lineups

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 U.S. 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 – either 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 imperil the identified individual, 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 U.S., having played a role in over 70% of the 330 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 (a) eyewitness confidence is an unreliable indicator of accuracy (3,4) and (b) 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 U.S. 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., 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 that 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.

Significance

In contrast to prior research, recent studies of simulated crimes have reported that (a) eyewitness confidence can be a strong indicator of accuracy and (b) 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 simultaneous 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.

Reserved for Publication Footnotes

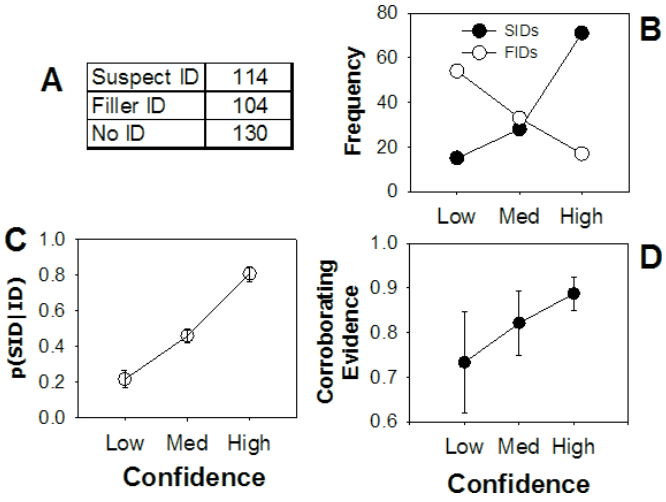

Fig. 1. (**A**) Frequency counts of eyewitness decisions in the Houston field study for 187 blind simultaneous and 161 blind sequential lineups combined. (**B**) Frequency of Suspect IDs (SIDs) and Filler IDs (FIDs) in **A** exhibited opposite trends as a function of confidence (low, medium or high), $\chi^2(2) = 55.3$, $p < .0001$. (**C**) For IDs made to a suspect or filler, the probability that it was a suspect ID 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 one-tailed Cochran–Armitage trend test, the effect was marginally significant, $Z = 1.57$, $p = .055$. Error bars represent standard errors.

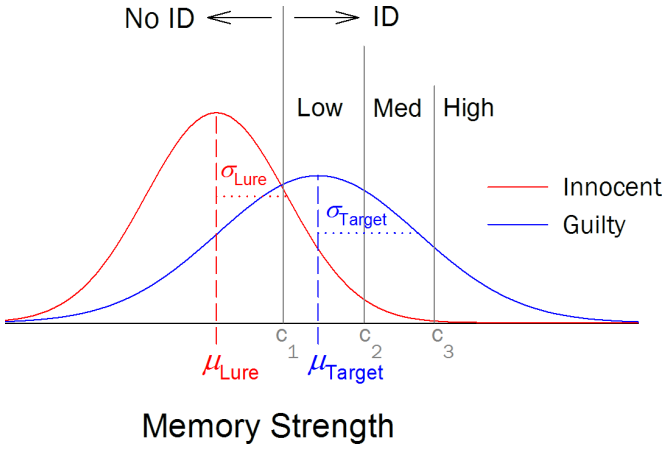

Fig. 2. Signal-detection conceptualization of low, medium or high confidence ratings associated with a positive ID. Memory strength (i.e., familiarity) values for lures (innocent suspects and fillers combined for a fair lineup) and for targets (guilty suspects) are distributed according to Gaussian distributions (red = lures, blue = targets) with means of μ_{Lure} and μ_{Target}, respectively, and standard deviations of σ_{Lure} and σ_{Target}, respectively. A fair 6-member target-present lineup is conceptualized as 5 random draws from the lure distribution and 1 random draw from the target distribution, and a fair 6-member target-absent lineup is conceptualized as 6 random draws from the lure distribution. Using the simplest decision rule, an ID is made if the most familiar person in a lineup exceeds c1, with confidence (Low, Medium or High) being determined by the highest criterion that is exceeded. With μ_{Lure} and σ_{Lure} set to 0 and 1, respectively, the model has 5 parameters (μ_{Target}, σ_{Target}, *c1*, *c2*, and *c3*), all scaled in units of σ_{Lure}. When fit to data produced by many participants, the model conceptualizes group performance (not the performance of any single participant). An equal-variance version of the model ($\sigma_{Lure} = \sigma_{Target}$), which eliminates one parameter, allows the addition of a base-rate parameter (p_{Target}) that can be used to estimate the proportion of target-present lineups in data that have been aggregated across target-present and target-absent lineups (as police department field data necessarily are).

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 (a) used photo lineups pseudo-randomly assigned to simultaneous (n = 187) or sequential (n = 161) formats, (b) were administered by an investigator who was blind to the identity of the suspect, and (c) involved suspects who were strangers to the eyewitnesses. Eyewitnesses who made suspect IDs or filler IDs from these lineups were asked to supply a confidence rating using a 3-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 novel 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, 6-person lineup, the suspect should be identified by a mock witness only 1/6 (.17) of the time. The mean proportion of suspect IDs made by the mock witnesses (.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 identifications collapsed across lineup format (i.e., simultaneous and sequential data combined). 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 five 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. 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), $\chi^2(1) = 31.02$, $p < .0001$, 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 5-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 sus-

A

Lineup	Decision	Frequency
TP (n = 456)	Guilty Suspect ID	253
	Filler ID	98
	No ID	105
TA (n = 452)	Incorrect ID	187
	Innocent Suspect ID	~23.4
	Filler ID	~163.6
	No ID	265

B

Suspect ID	276.4
Filler ID	261.6
No ID	370

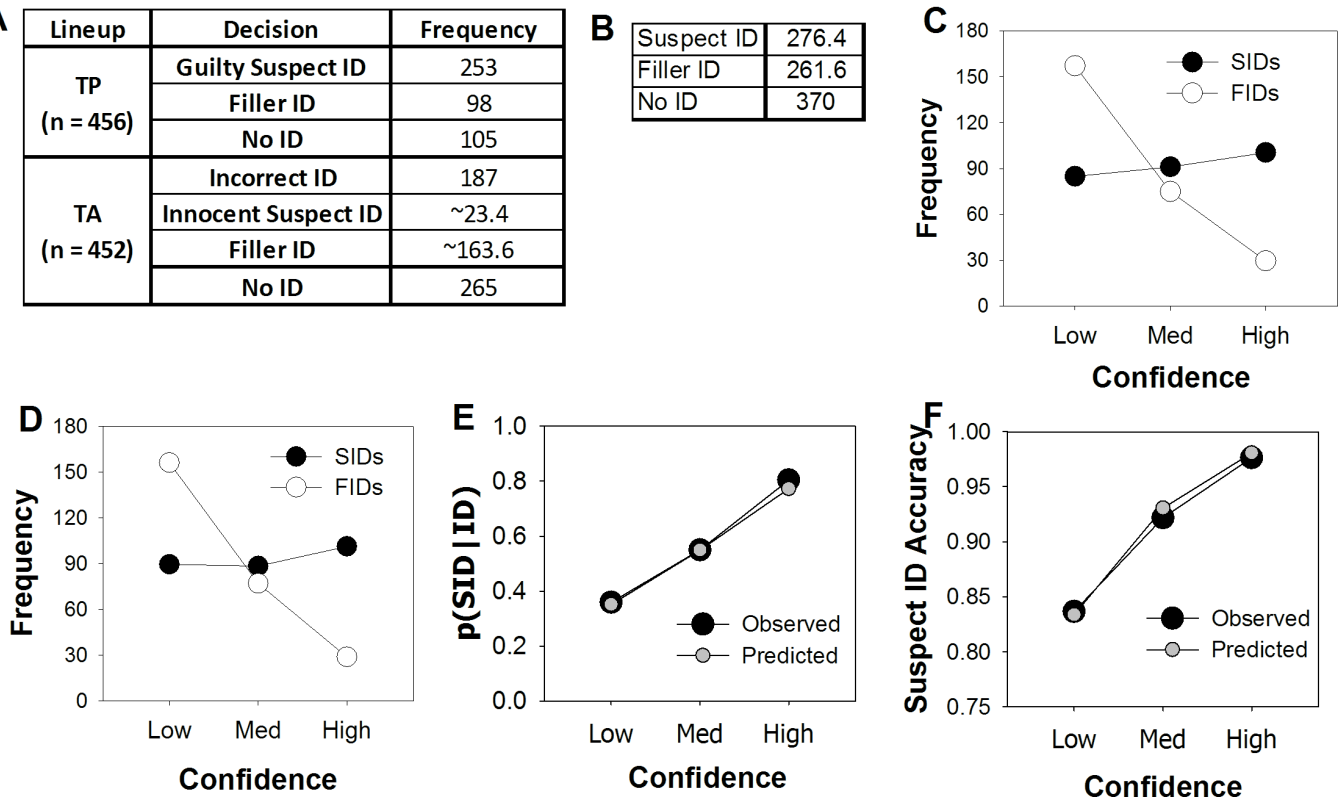

Fig. 3. (**A**) Frequency counts of eyewitness decision outcomes for 456 target-present (TP) and 452 target-absent (TA) simultaneous lineups from an experimentally-controlled field study (11). The TA lineups did not have a designated "innocent suspect." Therefore, following standard practice, an estimate ($\sim$) of the frequency of innocent suspect IDs from TA lineups was obtained by dividing all incorrect TA IDs (n = 187) by the lineup size of 8, with the remainder of incorrect IDs providing an estimate of the frequency of filler IDs. (**B**) Eyewitness decision outcomes in **A** summed (i.e., collapsed) across TP and TA lineups. (**C**) Frequency of Suspect IDs (SIDs) and Filler IDs (FIDs) in **B** as a function of confidence (low, medium or high). For this plot, the 100-point confidence scale was reduced to a 3-point scale (90-100 = High, 70-80 = Medium, and 0-60 = Low). (**D**) Predicted frequency of Suspect IDs, Filler IDs and No IDs based on a fit of the signal-detection model (Fig. 2) to the data in **C**. The fit was very good, $\chi^2(2) = 2.68$. (**E**) The observed proportion of IDs in **C** that were suspect IDs (black symbols) increased dramatically with confidence, as did the predicted values (small gray symbols) computed from the predicted values in **D**. (**F**) The proportion of suspect IDs in **C** that were guilty suspect IDs (black symbols) also increased dramatically with confidence, an effect that was accurately predicted by the signal-detection model (small gray symbols) despite its having been fit to data collapsed across TP and TA lineups.

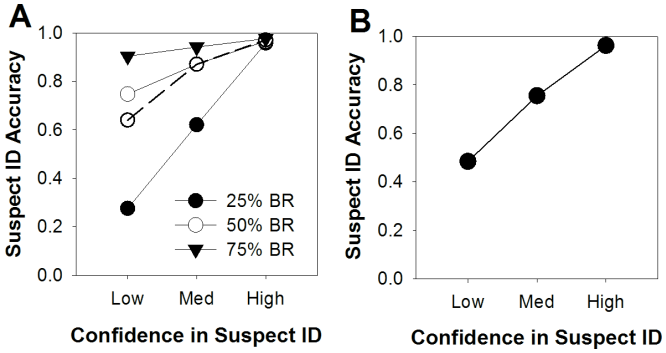

Fig. 4. (**A**) Signal-detection estimates of the posterior probability of guilt associated with suspects identified from lineups in the Houston field study for three different hypothetical base rates (BR). The estimates are averaged across simultaneous and sequential lineups. The dashed line shows the estimates from the high-threshold model assuming a 50% base rate. (**B**) Model-based estimate of the posterior probability of guilt associated with suspects identified from lineups in the Houston field study assuming an equal-variance signal-detection model (as suggested by fits to the experimentally-controlled field data) and including "target-present base rate" as a free parameter (estimated to be .35).

pect 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, of 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 witnesses, $1 - p$, some proportion of them, g, will make a random identification from the lineup despite not recognizing the perpetrator. For a fair, 6-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, $(1\text{-}p) * g * (1/6)$. 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, nS_{TP}:

$nS_{TP} = n_{TP} [\, p + (1\text{-}p) * g * (1/6) \,]$ **[Equation 1]**

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, $(1\text{-}p) * g * (5/6)$. Thus, the number of filler IDs from target-present lineups, nF_{TP} is:

$nF_{TP} = n_{TP} [\, (1\text{-}p) * g * (5/6) \,]$ **[Equation 2]**

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 $g * (1/6)$, and the probability of a filler ID from a fair target-absent lineup is $g * (5/6)$. 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:

$nS_{TA} = n_{TA} * [\, g * (1/6) \,]$ **[Equation 3]**

$nF_{TA} = n_{TA} * [\, g * (5/6) \,]$ **[Equation 4]**

These equations underscore the important fact that, for fair lineups, incorrect suspect IDs should be relatively rare compared to 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 (Equation 1 + Equation 3) and F is equal to sum of filler IDs from target-present and target-absent lineups (Equation 2 + Equation 4):

$S = n_{TP} [\, p + (1\text{-}p) * g * (1/6) \,] + n_{TA} * [\, g * (1/6) \,]$

$F = n_{TP} [\, (1\text{-}p) * g * (5/6) \,] + n_{TA} * [\, g * (5/6) \,]$

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:

$g = (6 * F) / (10 * n - 5 * S + F)$ **[Equation 5]**

$p = (5 * S - F) / (5 * n)$ **[Equation 6]**

Note that, using Equations 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, Equations 1 and 3 can now be used to estimate nS_{TP} and nS_{TA}, which can then be used to compute suspect ID accuracy, S_{acc}:

$S_{acc} = nS_{TP} / (nS_{TP} + nS_{TA})$ **[Equation 7]**

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 Equations 5 and 6, $g = .49$ and $p = .54$. Using these parameters, Equations 1 and 3 indicate that $nS_{TP} = 99.8$ and $nS_{TA} = 14.2$, so overall suspect ID accuracy (Equation 7) comes to $99.8 / (99.8 + 14.2) = .88$ (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 .97, .87 and .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 .62 for simultaneous lineups and .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 8-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 to 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 chi-square 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 to the collapsed observed values by computing a chi-square 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 chi-square statistic, yielding the final predicted values in Fig. 3D. An equal-variance model

turned out to be adequate (i.e., σ_{Target} did not differ significantly from 1, thus $\sigma_{Target} = \sigma_{Lure}$). 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).

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 σ_{Target} and σ_{Lure} to differ. The model was fit to the simultaneous and sequential data separately and also to the data combined across lineup format (*Table S4*). 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_{acc}) for the Houston field data – that is, it shows estimated values of $nS_{TP} / (nS_{TP} + nS_{TA})$ – 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 ($\sigma_{Target} = \sigma_{Lure}$) also applies to the Houston field data. Removing the unequal-variance parameter made it possible to add a base-rate parameter (p_{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 .35 for both lineup formats. That is, assuming the equal-variance model is correct, 35% of the photo lineups contained a guilty suspect and 65% contained an innocent suspect. At first glance, this relatively low estimate of the proportion of lineups containing a guilty suspect might be regarded as problematic. However, the confidence-accuracy relationship predicted by this best-fitting model (averaged across the predictions made by separate fits to the simultaneous and sequential data) exhibits a strong relationship between the confidence associated with a suspect ID and the accuracy of that ID (Fig. 4B). In other words, high-confidence IDs are accurate despite the low base rate of target-present lineups.

Simultaneous vs. Sequential Lineups. We also analyzed the data separately for simultaneous and sequential lineups (*Table S1*), focusing first on corroborating evidence of guilt associated with identified suspects. In a previous police department field study conducted in Austin, Texas, expert ratings of corroborating evidence of guilt suggested that innocent suspects were less likely to be identified and guilty suspects were more likely to be identified from simultaneous lineups than sequential lineups (22). Similarly, in the current Houston police department field study, the proportions of suspects identified from simultaneous (SIM) lineups (n = 68) and (SEQ) sequential lineups (n = 46) rated by the investigating officer as having independent evidence of guilt against them were SIM = .912 and SEQ = .761, $\chi^2(1) = 4.92$, $p = .027$. That is, according to this proxy measure of guilt, more of the suspects identified from simultaneous lineups were likely to be guilty – and fewer innocent – than suspects identified from sequential lineups. However, on a post ID questionnaire, the investigating officer noted that some of these witnesses (n = 65) reported that they (a) encountered a photo of the suspect before being presented with the photo lineup, (b) were under the influence of alcohol when they witnessed the crime, and/or (c) were not wearing their prescribed glasses during the crime (*Table S2*). The reported differences on the three questionnaire measures, if they were true and had any effect, would have worked against the sequential procedure. Yet 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 = .920 vs. SEQ = .789), $\chi^2(1) = 3.12$, $p = .077$. Thus, eliminating these 65 eyewitnesses reduced statistical power without having an appreciable effect on the pattern of results.

As indicated earlier, a 5-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 = .912 vs. SEQ = .804), $\chi^2(1) = 2.77$, $p = .096$. However, when the reduced recoded data set was analyzed (eliminating 65 witnesses based on their questionnaire responses), the effect, while continuing to favor the simultaneous procedure (SIM = .920, SEQ = .842), was no longer marginally significant, $\chi^2(1) = 1.30$, $p = .244$. Although not significant, even for this analysis, more suspects identified from simultaneous lineups had independent corroborating evidence of guilt compared to 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 p_{Target} fixed at .35 (free parameters = μ_{Target}, *c1, c2,* and *c3*), separately to the simultaneous and sequential Houston field data broken down by confidence (*Table S1*). When the full data set was analyzed, μ_{Target} was significantly higher for the simultaneous procedure than the sequential procedure (SIM = 2.87 vs. SEQ = 2.06), $\chi^2(1) = 4.86$, $p < .03$. When the reduced data set was analyzed (excluding the 65 witnesses discussed above), the difference in the estimated value of μ_{Target} still favored the

simultaneous procedure (SIM = 2.67 vs. SEQ = 2.13), but the effect was no longer significant, $\chi^2(1) = 2.51$, $p = .11$. 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 lab-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 non-crime 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. 4A). In addition, and again contrary to a widely-held view, the present results reinforce both ROC analyses of lab-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 (though, 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, non-blind 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 (a) the robberies involved strangers, and (b) 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 pseudo-randomly 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 6 photos (one suspect and 5 fillers). For the simultaneous procedure, the eyewitness viewed all 6 photos at the same time. For the sequential procedure, the 6 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 3-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 have played an instrumental role in the experiment: Executive Assistant Chief Martha Montalvo, Assistant Chief George Buenik, Captain Mark Holloway, Captain Lori Bender, Captain Heather Morris, 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 No. SES-1456571 to John T. Wixted.

1. Police Executive Research Forum (2013) *A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies*. Retrieved from http://www.policeforum.org/.
2. Innocence Project (2015) http://www.innocenceproject.org/causes-wrongful-conviction Downloaded July 28, 2015.
3. Lacy JW, Stark CEL (2013) The neuroscience of memory: implications for the courtroom. *Nat. Rev. Neurosci.* **14**, 1-10.
4. Penrod S, Cutler B (1995) Witness confidence and witness accuracy: Assessing their forensic relation. *Psychol Public Pol L* **1**, 817–845.
5. Lindsay RCL, & Wells GL (1985) Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *J Appl Psychol*, **70**, 556-564.
6. Steblay NK, Dysart JE, Fulero S, Lindsay RCL (2001) Eyewitness accuracy rates in sequential and simultaneous lineup presentations: a meta-analytic comparison. *Law Human Beh* **25**, 459-473.
7. Steblay NK, Dysart JE, Wells GL (2011) Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychol Public Pol Law* **17**, 99-139.
8. New Jersey Judiciary, *Supreme Court Releases Eyewitness Identification Criteria for Criminal Cases*, *July 19, 2012*, available at: http://www.judiciary.state.nj.us/pressrel/2012/pr12071-9a.htm.
9. Brewer N, Wells GL (2006) The confidence-accuracy relation in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *J Exp Psy: Applied* **12**, 11-30.
10. Sauer J, Brewer N, Zweck T, Weber N (2010) The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law Human Beh* **34**, 337-347.
11. Palmer M, Brewer N, Weber N, Nagesh A (2013) The confidence-accuracy relationship for eyewitness identification decisions: effects of exposure duration, retention interval, and divided attention. *J. Exp. Psychol.: App.* **19**, 55-71.
12. Wixted JT, Mickes L, Clark SE, Gronlund SD, Roediger HL (in press) Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *Am. Psychol.*.
13. Behrman BW, Davey SL (2001) Eyewitness identification in actual criminal cases: An archival analysis. *Law Human Beh* **25**, 475-491.
14. Behrman BW, Richards RE (2005) Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law Human Beh* **29**, 279-301.
15. Lusted LB (1971) Signal-detectability and medical decision-making. *Science* **171**, 1217-1219.
16. Swets JA (1979) ROC analysis applied to the evaluation of medical imaging techniques. *Invest Radiol* **14**, 109-121.
17. Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.
18. Gronlund SD, Carlson CA, Neuschatz JS, Goodsell CA, Wetmore SA, Wooten A, Graham M (2012) Showups versus lineups: An evaluation using ROC analysis. *J Appl Res Mem Cognition* **1**, 221-228.
19. Mickes L, Flowe HD, Wixted JT (2012) Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *J Exp Psychol - Appl* **18**, 361-376.
20. Carlson CA, Carlson MA (2014) An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *J Appl Res Mem Cognition*, **3**, 45-53.
21. Dobolyi DG, Dodson CS, (2013) Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *J Exp Psychol - Appl* **19**, 345-357.
22. Amendola KL, Wixted JT (2015) Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *J Exp Crim*. **11**, 263-284.
23. Wells GL, Steblay NK, Dysart JE (2015) Double-blind photo-lineups using actual eyewitnesses: an experimental test of a sequential versus simultaneous lineup procedure. *Law Human Beh*. **39**, 1-14.
24. Wells W (2014) The Houston Police Department eyewitness identification experiment: analysis and results. Retrieved from: http://www.lemitonline.org/research/projects.html
25. National Research Council (2014) *Identifying the culprit: Assessing eyewitness identification*. Washington, DC, National Academies Press.
26. Macmillan NA, Creelman CD (2005) *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
27. Dunn JC (2004) Remember–know: A matter of confidence. *Psychol. Rev.* **111**, 524–542.
28. Wixted JT (2007) Dual-process theory and signal-detection theory of recognition memory. *Psychol Rev* **114**, 152-176.
29. Wells GL, Bradfield AL (1998) "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *J Appl Psychol* **83**, 360-376.