Comparing the Diagnostic Accuracy of Suspect Identifications made by Actual Eyewitnesses from Simultaneous and Sequential Lineups in a Randomized Field Trial

Karen L. Amendola[1] & John T. Wixted[2]

[1]Police Foundation

[2]University of California, San Diego

*Correspondence concerning this article should be addressed to Karen L. Amendola, Police Foundation, 1201 Connecticut Avenue, NW, Suite 200, Washington, DC 20036-2636. E-mail: kamendola@policefoundation.org or John Wixted, Department of Psychology, 0109, University of California, San Diego, La Jolla, CA 92093-0109. E-mail: jwixted@ucsd.edu.

Abstract

Objectives: Eyewitness misidentifications have been implicated in many of the DNA exoneration cases that have come to light in recent years. One reform designed to address this problem involves switching from simultaneous lineups to sequential lineups, and our goal was to test the diagnostic accuracy of these two procedures using actual eyewitnesses.

Methods: In a recent randomized field trial comparing the performance of simultaneous and sequential lineups in the real world, suspect ID rates were found to be similar for the two procedures. Filler ID rates were found to be slightly (but, in the key test, nonsignificantly) higher for simultaneous than sequential lineups, but fillers will not be prosecuted even if identified. Moreover, filler IDs may not provide reliable information about innocent suspect IDs. Here, we use two different proxy measures for ground truth of guilt vs. innocence for *suspects* identified from simultaneous or sequential lineups in that same field study.

Results: The results indicate that innocent suspects are, if anything, *less* likely to be mistakenly identified – and guilty suspects are more likely to be correctly identified – from simultaneous lineups compared to sequential lineups.

Conclusions: Filler identifications are not necessarily predictive of the more consequential error of misidentifying an innocent suspect. With regard to actual suspect identifications, simultaneous lineups are diagnostically superior to sequential lineups. These findings are consistent with recent lab-based studies using receiver operating characteristic analysis suggesting that simultaneous lineups make it easier for eyewitnesses to tell the difference between innocent and guilty suspects.

Keywords: Eyewitness Identification, ROC Analysis, Sequential Lineups, Simultaneous Lineups

Comparing the Diagnostic Accuracy of Suspect Identifications made by Actual Eyewitnesses from Simultaneous and Sequential Lineups

More than 300 people have been exonerated by DNA evidence in recent years, and many of those individuals were wrongfully convicted, at least in part, based on eyewitness misidentifications. The apparent unreliability of eyewitness identification evidence has motivated a concerted effort to find some way to reduce this problem, and much of the focus in this regard has been placed on trying to determine whether sequential lineups should replace simultaneous lineups. Recently, these two lineup procedures were compared using real eyewitnesses in a study known as the American Judicature Society (AJS) field study. Phase 1 results from that study (Wells, Steblay & Dysart, 2011, 2014) focused on the proportion of simultaneous and sequential lineups associated with suspect IDs, filler IDs, and lineup rejections. The proportion of suspect IDs was similar for the two procedures (25% for simultaneous lineups and 27% for sequential lineups), but filler IDs were higher for the simultaneous procedure (18% for simultaneous lineups vs. 12% for sequential lineups). Although the difference in filler ID rates was not statistically reliable when based on the final decisions made by eyewitnesses in the sequential procedure[1], Wells et al. (2014) nevertheless attached interpretative significance to this non-significant effect. Specifically, because fillers are known to be innocent, the authors of the study inferred that innocent suspects are also more likely to be incorrectly identified from simultaneous lineups than from sequential lineups. Here, we report Phase 2 results focusing on measures of likely guilt associated with the suspects who were identified from simultaneous and sequential lineups in the AJS field study. Because suspect IDs – especially *innocent* suspect IDs – are far more consequential than filler IDs, this approach more directly addresses the question of whether

simultaneous or sequential lineups lead to fewer false IDs of the innocent and more correct IDs of the guilty.

Background

In the simultaneous procedure, the members of the lineup (usually 6 people – 1 suspect and 5 fillers) are presented together, whereas in the sequential procedure, the members of the lineup are presented one at a time for individual recognition decisions. Many mock-crime laboratory studies have evaluated the performance of these two lineup procedures to determine if sequential lineups lead to fewer false IDs of innocent suspects than simultaneous lineups and, more generally, to determine if sequential lineups are diagnostically superior to simultaneous lineups. In these lab studies, some participants view a lineup in which the suspect is, in fact, the perpetrator (target-present lineups), but other participants view a lineup in which the suspect is an innocent person who resembles the perpetrator (target-absent lineups). The proportion of target-present lineups from which the guilty suspect is correctly identified is called the correct ID rate, and the proportion of target-absent lineups from which the innocent suspect is incorrectly identified is called the false ID rate. Ideally, one would like to maximize the correct ID rate and minimize the false ID rate. Because the fillers in a lineup are not suspects and are therefore known to be innocent, a filler ID does not endanger the identified individual and is therefore not treated as the equivalent of a false ID.

In a recent meta-analysis, Steblay, Dysart and Wells (2011) found that the average correct and false ID rates for the simultaneous lineup procedure (computed without regard for filler IDs) were 0.52 and 0.28, respectively, whereas the corresponding values for the sequential lineup procedure were 0.44 and .15, respectively[2]. This outcome appears to favor the sequential procedure because the decrease in the false ID rate (from .28 to .15) considerably exceeds the

decrease in the correct ID rate (from .52 to .44). Intuitively, the cost (namely, the small decrease in the correct ID rate) seems worth the benefit (namely, the large decrease in the false ID rate).

The performance of the two lineup procedures is often summarized by a single measure known as the diagnosticity ratio, which is equal to the correct ID rate divided by the false ID rate. Steblay et al. (2011) found that the diagnosticity ratio was higher for the sequential procedure ($0.44 \div 0.15 = 2.93$) than the simultaneous lineup procedure ($0.52 \div 0.28 = 1.86$). A higher diagnosticity ratio implies higher *posterior odds of guilt* (which are the odds that a suspect who has been identified from a lineup is actually guilty). Thus, according to the data analyzed by Steblay et al. (2011), a suspect identified from a sequential lineup is more likely to be guilty than a suspect identified from a simultaneous lineup. On the surface, the overall case in favor of the sequential lineup seems compelling because (one might assume) switching to the sequential procedure in the real world would lower the false ID rate while increasing the trustworthiness of a suspect ID.

Intuition notwithstanding, findings like these do not indicate that sequential lineups are diagnostically superior to simultaneous lineups, nor do they suggest that switching to sequential lineups in the real world would reduce the frequency of false IDs. In fact, sequential lineups might *reduce* diagnostic accuracy and *increase* the risk to innocent suspects even if the findings analyzed by Steblay et al. (2011) are accurate. Many researchers do not accept their interpretation of the literature as being accurate (e.g., Clark, 2012; Gronlund, Carlson, Dailey & Goodsell, 2009; McQuiston-Surrett, Malpass & Tredoux, 2006) but disputing their interpretation is not our purpose here.

A non-intuitive fact that has only recently been taken into consideration by the field is that the diagnostic performance of a given lineup procedure cannot be adequately characterized

by a single correct and false ID rate pair but can only be adequately characterized by an entire family of correct and false ID rates (Gronlund, Mickes & Wixted, 2014; Wixted & Mickes, 2012). Perhaps the easiest way to appreciate the fact that more than one correct and false ID rate characterizes a given lineup procedure is to consider two otherwise identical jurisdictions that differ in only one respect: Jurisdiction A includes a "not sure" response option when eyewitnesses are presented with a simultaneous lineup, whereas Jurisdiction B does not. In Jurisdiction A, eyewitnesses who are not confident of their ability to identify the perpetrator from the lineup would sometimes choose the "not sure" response option instead of making a low-confidence ID. In Jurisdiction B, eyewitnesses who are not confident of their ability to identify the perpetrator from the lineup – and who would choose the "not sure" response option if it were available – would make a low-confidence ID instead. Because more IDs (correct and incorrect) would be observed in Jurisdiction B than in Jurisdiction A, the correct and false ID rates would be higher in Jurisdiction B compared to Jurisdiction A. In that case, there would be *two* sets of correct and false ID rates for the simultaneous lineup, and neither one would be more valid than the other. If, in addition to including a "not sure" response option, Jurisdiction C also included an explicit instruction informing eyewitnesses that they do not have to choose anyone from the lineup (further reducing the pressure to choose), the correct and false ID rates in that jurisdiction might be even lower than those observed in Jurisdictions A or B. This third pair of correct and false ID rates for the simultaneous procedure is as valid as the other two.

The key point is that a lineup procedure (whether simultaneous or sequential) is characterized by an entire family of correct and false ID rates obtained by adjusting the overall tendency of eyewitnesses to make an ID from the lineup – a tendency that policymakers can manipulate (e.g., by including a "not sure" response option and/or by including instructions that

reduce the pressure an eyewitness might feel to make an ID). A variable that policymakers can manipulate is known as a *system variable* (Wells, 1978). The fact that lineup instructions can be used to reduce the pressure an eyewitness might feel to choose (i.e., to induce a more conservative decision criterion) has been noted before (Clark, 2005; Brewer, Weber & Semmler, 2005), but the implications of that fact have rarely been considered. The implications are more important than they might seem to be at first glance.

If a given lineup procedure (e.g., the simultaneous procedure) is characterized by more than one correct and false ID rate, it follows that it is also characterized by more than one diagnosticity ratio. That being the case, it can be misleading to compare a singular diagnosticity ratio for the simultaneous procedure (by choosing one from its family of diagnosticity ratios) to a singular diagnosticity ratio for the sequential procedure (by choosing one from its family of diagnosticity ratios). In particular, it is misleading when overall suspect choosing rates differ for the two procedures being compared (Wixted & Mickes, 2012), as they usually do for simultaneous and sequential lineups. For example, as noted above, Steblay et al. (2011) found that suspect choosing rates – both the correct ID rate and the false ID rate – were relatively high for the simultaneous lineup procedure (average correct and false ID rates were 0.52 and 0.28, respectively) compared to the sequential lineup procedure (average correct and false ID rates were 0.44 and .15, respectively). When overall choosing rates differ like that, it is not meaningful to compare the diagnosticity ratios (or, equivalently, the posterior odds of guilt) because that measure increases dramatically as the choosing rate (i.e., the overall tendency of witnesses to make an ID) decreases for either procedure. Thus, the fact that a procedure with a lower choosing rate has a higher diagnosticity ratio is not, in itself, a particularly informative finding.

It might be tempting to ignore this technical argument about diagnosticity ratios and to concentrate instead on the large difference between the false ID rates associated with the two lineup procedures – a result that appears to suggest that innocent suspects are placed at much greater risk when simultaneous lineups are used compared to when sequential lineups are used. However, appearances can be misleading. For example, Wells, Steblay and Dysart (2012) argued that the extra correct and false IDs associated with the simultaneous procedure may result from random guesses, which are less likely to occur than when a sequential procedure is used. This possibility raises an interesting question: what would the correct and false ID rates be when low-confidence guesses are eliminated from consideration for both lineup procedures?

As noted above, one way to reduce the impact of random guesses would be to include a "not sure" response option, which allows witnesses to avoid making an ID by choosing that option instead of guessing. Under those conditions, the correct and false ID rates would both decrease. Imagine that the correct and false ID rates for the sequential procedure decrease to .40 and .10, respectively (down from .44 and .15, respectively), and the correct and false ID rates for the simultaneous procedure decrease to .45 and .05, respectively (down from .52 and .28, respectively). These new correct and false ID rates are purely hypothetical and were deliberately chosen to illustrate the possibility that, using the traditional metrics (i.e., the false ID rate and the diagnosticity ratio), simultaneous lineups could be superior to sequential lineups when the effects of guessing are minimized. In this hypothetical example, the simultaneous lineup has both a lower false ID rate (.10 for sequential; .05 for simultaneous) and a higher diagnosticity ratio (.40 ÷ .10 = 4 for sequential; .45 ÷ .05 = 9 for simultaneous).

Which correct and false ID rate pair should be used to decide whether or not one procedure is superior to other? The first pair that included guesses or the second (more

conservative) pair that excluded guesses? Considerations like these illustrate why receiver-operating characteristic (ROC) is needed to evaluate the diagnostic accuracy of competing lineup procedures. ROC analysis involves nothing more than examining the full range of correct and false ID rates that arise for a single lineup procedure as the tendency to identify someone from the lineup varies over a wide range (while holding *discriminability* – which is the ability to tell the difference between an innocent suspect and a guilty suspect – constant). The ROC analytic method was first developed in World War II by mathematicians and engineers seeking better ways to measure the diagnostic performance of radar and sonar, but it is now widely used in many applied fields, including diagnostic medicine. Previously published articles provide a detailed introduction to ROC analysis in the eyewitness domain, explaining how to do it, why it is necessary, and why it is the method of choice in many other applied fields (Gronlund, Mickes & Wixted, 2014; Wixted & Mickes, 2012).

Recent ROC analyses have consistently found that the simultaneous lineup yields a higher ROC – that is, the simultaneous lineup yields higher diagnostic accuracy – than the sequential lineup (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes, Flowe & Wixted, 2012). What does this result actually mean? First, it means that simultaneous lineups make it easier for eyewitnesses to tell the difference between innocent and guilty suspects. Second, and critically, it means that if suspect choosing rates happened to be the same for simultaneous and sequential lineups, then it would have to be the case that the false ID rate would be lower and the correct ID rate would be higher for the *simultaneous* procedure.

When choosing rates are the same (as they were in the AJS field study, Phase I), one can simply refer to the correct and false ID rates to easily determine which procedure is superior, as in the hypothetical example presented above. For the sequential lineup, the correct and false ID

rates in that example were chosen to be .40 and .10, respectively. For the simultaneous lineup, the corresponding values were .45 and .05. Thus, the overall suspect choosing rate[3] for the sequential lineup is (.40 + .10) ÷ 2 = .25, and the overall suspect choosing rate for the simultaneous lineup is the same, namely, (.45 + .05) ÷ 2 = .25. When the choosing rates are the same, the correct and false ID rates clearly indicate which procedure is superior (the simultaneous procedure in this example). But one can also use the diagnosticity ratio, or the posterior of odds of guilt, to make that determination. These measures are problematic when suspect choosing rates differ for the two procedures (because their values increase when the choosing rate is reduced by inducing more conservative responding for either procedure), but when choosing rates are the same, a measure like the posterior odds of guilt can be used to directly identify the superior procedure. In this example, the posterior odds of guilt are higher for the simultaneous procedure (.45 ÷.05 = 9) than for the sequential procedure (.40 ÷.10 = 4). This means that a suspect identified from a simultaneous lineup is 9 times more likely to be guilty than innocent, whereas a suspect identified from a sequential lineup is only 4 times more likely to be guilty than innocent. The performance of the two lineup procedures can also be quantified using the posterior probability of guilt, which in this example is higher for the simultaneous procedure (.45 / [.45 + .05] = .90) than the sequential procedure (.40 / [.40 + .10] = .80).

The critical point of this hypothetical example is that if suspect choosing rates happen to be the same for both lineup procedures, as they were in the AJS field study, then the posterior probability of guilt for suspects identified from each procedure would unambiguously indicate which procedure is diagnostically superior. Specifically, the procedure associated with the higher posterior probability of guilt would necessarily have both a higher correct ID rate and a lower

false ID rate than the other procedure. This raises a key question: Which procedure yielded the higher posterior probability of guilt in the AJS field study?

Measuring the posterior probability of guilt requires information about the ground truth of the guilt or innocence of identified suspects, and that information is usually not available in a field study. Indeed, this is precisely why Wells et al. (2011, 2014) relied on filler IDs as a proxy for the false ID rate. However, in our analysis of the data generated in Phase 2 of the AJS field study, we used case dispositions (Study A) and expert ratings (Study B) as proxies for the ground truth of guilt vs. innocence. Our goal was to estimate the posterior probability of guilt for suspects who were identified from simultaneous and sequential lineups in the AJS field study.

The AJS Field Study

In response to calls for a more robust field study, the American Judicature Society implemented a randomized field trial designed to compare sequential and simultaneous presentation methods in multiple field sites (Wells, et al., 2011). Wells et al. (2011, 2014) implemented that experiment in four sites: Charlotte-Mecklenburg County, North Carolina; Tucson, Arizona; San Diego, California; and Austin (Travis County), Texas. In this study, all factors other than the presentation method were held constant. The protocol required standardized instructions administered via a laptop presentation mode and ensured that all lineup administrations were double blind. The lineup presentation method itself – sequential versus simultaneous – was randomly assigned by computer for each witness immediately prior to viewing.

The data set consisted of 494 double-blind lineups from witnesses who were attempting to identify a suspect who was a stranger and who were seeing the suspect's photo for the first time. In laboratory studies, witnesses are usually told that the perpetrator may or may not be in

the lineup, and this instruction was included in the AJS field study as well. Eyewitnesses also were told that they would view all the individuals in the sequential lineup, and they were allowed to view the lineup a second time if requested. Critically, witnesses in the field study (unlike in the typical lab study) were given a "not sure" response option. This allowed witnesses to say that they were not sure, in which case they made no identification at all. The use of a "not sure" response option is conceptually similar to using a lineup instruction to induce more conservative responding, such as an instruction that says "Do not identify someone from the lineup if you are not sure of your decision." A few lab studies have found that providing eyewitnesses with an explicit “don’t know” option reduces suspect IDs (i.e., it leads to more conservative responding), yielding the expected increase in the diagnosticity ratio that generally accompanies more conservative responding (Perfect & Weber, 2012; Weber & Perfect, 2012; see also Steblay & Philips, 2011). In addition, to further reduce the pressure to choose, witnesses in the AJS field study were told that they "did not have to make an identification" and that "the investigation would continue even if they did not identify someone." These various methods (the "not sure" response option and special instructions designed to reduce the pressure to choose) would be expected to induce conservative responding and likely account for why Wells et al. (2011, 2014) found that in the AJS field study, overall suspect choosing rates were lower than the rates observed in previous studies.

As noted earlier, the considerations discussed above indicate that the suspect choosing rate is, to a certain degree, a system variable (i.e., it is under the control of the legal system), which means, for example, that the suspect choosing rate for simultaneous lineups could easily be reduced (e.g., by including a "not sure" response option, as was done in the AJS field study) if policymakers decided that the cost in terms of reduced correct IDs is worth the benefit in terms

of reduced false IDs. This point is important to appreciate because many are under the mistaken impression that simultaneous lineups are inferior to sequential lineups because simultaneous lineups yield higher correct and false ID rates. The key point is that switching to the sequential procedure is not the only way (and is not likely to be the best way) to lower suspect choosing rates. The methods used in the AJS field study illustrate another way to induce conservative responding, and when those methods are used, suspect choosing rates are reduced and turn out not to differ for simultaneous and sequential lineups. That fortuitous outcome created a unique opportunity to effectively evaluate the diagnostic accuracy of simultaneous and sequential lineups in the real world without having to perform ROC analysis.

Table 1 summarizes the most relevant results reported by Wells et al. (2011, 2014). For witnesses who requested a second viewing of the sequential lineup, their lap 2 decisions were used in this analysis because only those final decisions would be taken into consideration in a court of law. Wells et al. found that the two lineup procedures yielded similar suspect ID rates (25% for simultaneous and 27% for sequential, a negligible, non-significant difference), whereas filler ID rates differed to a greater degree (18% for simultaneous compared to 12% for sequential, though this was still not a significant difference, $p = .09$). For suspect and filler IDs combined, 44% of eyewitness made an ID from simultaneous lineups, and 40% of eyewitness made an ID from sequential lineups (also not a significant difference, $p > .35$). Thus, for these key results, there were no statistically reliable differences in the choosing rates for simultaneous and sequential lineups in the AJS field study.

As described earlier, when suspect ID rates are similar, the posterior probability of guilt provides an objective measure of which procedure has a lower false ID rate and a higher correct ID rate. In lab studies, the researcher knows which suspect IDs are correct and which are

incorrect, so the measure of interest (the diagnosticity ratio – that is, the posterior odds of guilt) can be directly computed. In the field study, the innocence or guilt of the suspect is not known. For that reason, Wells et al. (2011, 2014) used filler ID rates as a proxy measure. Because fillers are known to be innocent, Wells et al. reasoned that the procedure with the higher filler ID rate would also be the procedure with the higher innocent suspect ID rate. As they put it: "Hence, if the simultaneous procedure inflates rates of filler identifications relative to a sequential procedure, it logically follows that it also inflates risk to an innocent suspect" (p. 34).

In considering this claim, it should be kept in mind that the difference in simultaneous-vs.-sequential filler ID rates in the AJS field study was *not statistically significant* in the analysis of interest (i.e., in the analysis of final decisions, which included the lap 2 decisions made by witnesses who asked to view the sequential lineup a second time). Instead, the difference was significant only when it was based on lap 1 decisions (not taking into account the final decisions of witnesses who asked for a second viewing). Although that analysis is relevant to lab studies, which typically do not allow a second viewing, it is not relevant to how sequential lineups are typically used in actual practice, which is the analysis of interest to policymakers (i.e. the final decision by the witness/victim). It may not be prudent to attach interpretative significance to the non-significant difference in filler ID rates in the analysis of interest.

Moreover, even if the non-significant trend in filler ID rates is taken seriously, it is not necessarily true that filler ID rates serve as a valid proxy for innocent suspect ID rates. This point is most easily appreciated by considering the results from a lab study that were reported by Carlson, Gronlund, and Clark (2008). When the data from their Fair Condition are collapsed across target-present and target-absent lineups (as if it were a field study with suspect status unknown), the pattern of results looks very much like the pattern observed in the AJS field study.

Table 2 shows the collapsed data from Carlson et al. (2008). As in the AJS field study, overall filler choosing rates were higher for the simultaneous procedure (bolded values in the second row of data under "Collapsed"). However, unlike in the AJS field study, we can un-collapse these lab data to determine whether or not the overall filler choosing rate is a useful proxy for the innocent suspect choosing rate. Table 2 also presents those results (bolded values in the first row of data under "Target Absent"), and it is clear that, in this case, the sequential procedure yielded a *higher* (not a lower) innocent suspect ID rate, this despite the fact that the sequential procedure also yielded a lower filler identification rate. Thus, according to this study, filler ID rates do not necessarily predict innocent suspect ID rates (at least not when the data show the same pattern as was observed in the AJS field study). These findings serve as a reminder that intuitively reasonable inferences can be empirically wrong and therefore quite misleading.

In any case, the real question of interest has nothing to do with filler IDs (because fillers are “known innocents,” they are not endangered when identified by an eyewitness)[4] but instead has to do with the *ground truth* of guilt vs. innocence for suspects identified in the AJS field study. In our analysis of Phase 2 data, we focus specifically on measuring the ground truth regarding the guilt or innocence of suspects identified from simultaneous and sequential lineups in the AJS field trial. The key issue is whether the posterior probability of guilt is higher for one procedure or the other. Given that suspect choosing rates were similar, the procedure that yields the higher posterior probability of guilt is the one associated with a higher correct ID rate and a lower false ID rate. In Part A of our study, we track case outcomes across three of the four AJS field study sites (and ask: were the identified suspects ultimately adjudicated to be guilty or not guilty?) as a proxy measure of ground truth. In Part B of our study, we use expert ratings of evidentiary strength connecting the suspect to at least one of the crimes charged (as a proxy for

likely guilt) as assessed by actual police investigators, prosecutors, defense attorneys, and judges in Austin as a measure of ground truth.

Given the previous discussion, it is perhaps not surprising that the prediction derived from recent lab-based ROC analyses (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al, 2012; Mickes et. al., 2012) and the Wells et al. (2011, 2014) prediction derived from filler picks in the AJS field study are diametrically opposed. The ROC data indicate that simultaneous lineups are diagnostically superior to sequential lineups. In other words, in the lab, simultaneous lineups result in a higher number of guilty suspect IDs and fewer innocent suspect IDs than sequential lineups when the overall proportion of suspects identified from the two lineups is the same. Thus, the ROC-based prediction is that because the overall proportion of suspects identified from the two lineups was approximately the same in the AJS field study, the posterior probability of guilt (i.e., the probability that an identified suspect is guilty) will be higher for the simultaneous lineup than for the sequential lineup. This outcome would mean that the correct ID rate is higher, and the false ID rate is lower, for simultaneous lineups compared to sequential lineups. By contrast, using filler picks as a guide, the opposite prediction follows. Because the simultaneous procedure may inflate filler identifications relative to a sequential procedure, the prediction is that the simultaneous procedure also inflates the risk of misidentifying innocent suspects. In that case, the sequential procedure would be associated with a higher posterior probability of guilt. This outcome would mean that the correct ID rate is higher, and the false ID rate is lower, for sequential lineups compared to simultaneous lineups, which data to be presented here show to be untrue.

Study A: Analysis of Case Outcomes

What is the relationship between the lineup presentation method (sequential vs. simultaneous) and the case dispositions of identified suspects? If more innocent suspects were misidentified from simultaneous lineups than from sequential lineups (as might be assumed based on filler picks), then one would expect that a smaller proportion of suspects identified from simultaneous lineups would be found guilty. If, instead, more innocent suspects were identified from sequential lineups than from simultaneous lineups (as might be assumed based on recent ROC analyses conducted in the laboratory), then one would expect that a smaller proportion of suspects identified from sequential lineups would be found guilty.

Method

In order to ensure that the cases associated with the lineups from the AJS field study (Wells et al., 2011) had reached disposition, we required that at least one year pass since the lineups were presented. In order to assess the relationship between lineup presentation methods and case dispositions, we conducted an archival analysis with data collected from the AJS field study (Wells et al., 2011). We received disposition data from all four sites, and while the agencies were not able to provide us with dispositions for every case, we examined the data for all but one site. Because the descriptions of the outcomes varied by agency, we were only able to categorize the dispositions as having been adjudicated guilty (by plea or judgment) vs. not prosecuted. Dispositions from Charlotte-Mecklenburg County were not used because the study was prematurely discontinued based on changes in state law mandating the double-blind sequential procedure for lineup presentation. Thus, our analysis included cases from Austin, San Diego, and Tucson.

Results

The cases for which dispositions were reported by the agencies are presented in Table 3. As is shown in the Table, the rate of guilty judgments (by verdict or plea bargain) among these cases is 38%, with Austin having the highest (48%) as compared to just 25% in Tucson and 21% in San Diego. The rate of guilty judgments appears much lower than the national average of 78% in state courts, where the vast majority of all felony convictions in the U.S. occur (Durose & Langan, 2003). One possible explanation for the differences in conviction rates is that our data set primarily consisted of stranger crimes (suspect and victim unknown to each other), whereas in non-stranger crimes, the victim or witness often provides the name of the perpetrator and his/her relationship to the victim, rendering a lineup unnecessary. The other key reason is that more conservative criteria were used thereby lowering choosing rates (e.g. a "not sure" choice was made available; the instructions included both that "the suspect may or may not be in the lineup," and that "the investigation will continue whether or not you identify someone").

For present purposes, the key question concerns case dispositions for *suspects* identified from simultaneous and sequential lineups. We focus on suspect IDs because, with respect to lineups, the goal of the legal system is to maximize correct IDs (reducing the threat to society) while minimizing incorrect IDs (reducing the threat to innocent suspects). By comparison, filler IDs are relatively inconsequential because they do not increase or decrease the threat to anyone. Case disposition information was available for 32 suspects identified from a sequential lineup and 37 suspects identified from a simultaneous lineup.

What are the posterior odds of guilt for these suspect IDs? Of the 32 suspects identified from a sequential lineup, 21 were ultimately judged guilty and 11 were not prosecuted. Thus, by this measure, the posterior odds of guilt were $21 \div 11 = 1.91$. Of the 37 suspects identified from a simultaneous lineup, 26 were ultimately judged guilty and 11 were not prosecuted. Thus, by this

measure, the posterior odds of guilt were 26 ÷11 = 2.36. Expressed as a probability, the posterior probability of guilt for the sequential procedure, 21 / (21 + 11) = 0.656, was lower than the posterior probability of guilt for the simultaneous procedure, 26 / (26 + 11) = 0.703. Although the difference is small and not significant, the direction of the effect slightly favors the simultaneous lineup. Thus, these data offer no support for a sequential superiority effect in the real world and instead provide slight evidence for a simultaneous superiority effect (as predicted by recent lab-based ROC analyses).

It is important to emphasize that the finding by Wells et al. (2011, 2014) that simultaneous lineups lead to slightly more filler picks (a non-significant finding) ultimately *did not matter* in these cases in terms of the guilty or not prosecuted outcomes. This result indicates that "filler picks" are not necessarily representative of the more consequential error of picking an innocent suspect in a lineup. This conclusion accords with our earlier analysis of the Carlson et al. (2008) data summarized in Table 2. Based on the case disposition data we analyzed, 30% (11 out of 37) of suspects identified from a simultaneous lineup were not prosecuted (and were perhaps innocent), whereas 34% (11 out of 32) of suspects identified from a sequential lineup were not prosecuted (and were perhaps innocent). Thus, based on these results, if the goal is to protect innocent suspects, switching to the sequential lineup would not be advised.

Study B: Evidentiary Strength Study

Because the case disposition measure used in Part A may be a noisy measure of ground truth (e.g., case outcomes are partly determined by the skill of the attorneys involved), the present study also included a second and arguably much better proxy for ground truth, namely, an "evidentiary strength" scale developed in large part by a number of police investigators, defense attorneys, prosecutors, and judges under the guidance of Police Foundation researchers

(see Amendola & Slipka, 2009). The instrument uses a 5-point Likert scale where a "5" means that the evidence is particularly strong in linking to the identified suspect, and a "1" means that the evidence is exceptionally weak in linking to the identified suspect. The scale requires ratings across six categories of evidence (physical evidence, suspect statement information, suspect history, victim characteristics, witness characteristics, and identification information) plus an overall evidentiary strength rating. Exemplars are provided on the scale to give concrete illustrations of what a particular rating means. The case files for suspects identified from simultaneous and sequential lineups were rated by an expert team of decision makers in the criminal justice system (police investigators, prosecutors, defense attorneys, and judges) who were blind to the type of lineup that was used. One of the main questions of interest was whether suspects identified from simultaneous lineups had higher or lower ratings of guilt, on average, than suspects identified from sequential lineups.

Horry, Halford, Brewer and Milne (2014) argued that the use of corroborating evidence to establish the ground truth of guilt vs. innocence is potentially problematic if (1) the corroborating evidence influences police behavior (e.g., if it causes a non-blind lineup administrator to steer the witness towards the suspect) or (2) the eyewitness ID itself influences the search for further corroborating evidence. The first concern was minimized in the AJS field study by using blind administrators for both simultaneous and sequential lineups. The second concern, while valid, would presumably apply equally to simultaneous and sequential lineups and would therefore be unlikely to bias our findings in favor of one lineup procedure or the other.

Method

Site Selection

The study was conducted in Austin (Travis County), Texas, the site in the AJS field study (Wells et al., 2011) from which 70% of the data were generated. The three other sites were excluded from this site for a variety of reasons. First, two sites (Charlotte, NC and San Diego, CA) had limited sample sizes and the former had to discontinue participation early on when the state law mandated a sequential procedure. In Tucson, AZ a study had been underway for some time without District Attorney involvement in the AJS study, and prior to the establishment of a methodology for the outcome analysis. Another reason to focus on the Austin data was to minimize random error that might be introduced by site variance (e.g., error variance associated with differences in protocol adherence, or other characteristics of the respondents or agency culture).

Case Selection

The cases were initially selected from the overall pool of cases in the AJS field study in which all the experimental protocols had been followed in phase one (n= 340) and were thusly classified as "pristine" by Wells, et al. (2011). The cases included were criminal and primarily made up of assaults and aggravated assaults, burglaries, robberies, and thefts. Next, due to state law in Texas, and instructions from the District Attorney's Office, we also eliminated any cases involving juvenile suspects (n = 6) and lineups associated with cases that involved sexual assault (n = 6) resulting in 328 lineups that met the criteria of the agency and research team. Additionally, we eliminated the 15 cases that were referred to the county attorney's office (primarily due to their status as misdemeanors), resulting in a sample of 313 eligible lineups (156 simultaneous lineups and 157 sequential lineups).

A subset of these 313 cases was then randomly selected to be rated in the Phase 2 analysis. Specifically, we selected a random sample of 200 lineups[5] stratified by lineup

presentation method in order to obtain relative balance among the pick types. Note that this random sampling step was performed as part of a broader study (AJS field study Phase 2) which included an experimental study investigating the extent to which knowledge of a suspect ID or lineup procedure influenced the interpretation of evidentiary strength for other case evidence (see Amendola, et al., 2014). Here, we focus solely on evidentiary strength ratings associated with suspect identifications from simultaneous and sequential lineups because, as explained earlier, the probative value of these identifications directly indicates which lineup procedure is superior to the other. Upon further review of case details after the stratified random sampling procedure, an additional 49 lineups were found to be ineligible for inclusion by research staff (e.g., juvenile involvement, sexual assault, inconsistencies in case details, suspect not mentioned in case, etc.). After excluding these cases, the final analysis sample consisted of 151 cases (sequential n = 75; simultaneous n = 76). In this sample of cases, we had 22 suspect picks from a simultaneous lineup and 30 suspect picks from a sequential lineup to analyze. Filler picks were represented in 19 simultaneous lineups and 16 sequential lineups, and no picks were made in 29 of the sequentially presented lineups and 35 of the simultaneously presented lineups. These 151 photo arrays were rated by our team of case evaluators.

Participants

Case evaluators were selected from a recruited pool of 26 criminal justice decision makers (10 female and 16 male). The cases were rated in various sessions held in the fall of 2012. On a given day, cases were rated by 8 participants (two each of police investigators, prosecutors, defense attorneys, and judges). Some of the raters had career experience that fell into more than one category (e.g., two raters had prior experience serving as a district attorney, as a defense

attorney and as a judge) and could therefore serve in a different role on different days to balance out the expertise of the 8 raters.

Training

Training was provided to the participating criminal justice evaluators to explain how the instrument was developed, what the exemplars (rating scale anchors) represented, how they were derived, and how to rate each category of evidence independently. This training required a block of approximately four to five hours to complete.

Next, the evaluators practiced using the instrument on actual cases provided by an independent jurisdiction. This training began with a group session in which all of the case evaluators read the same case and came up with a rating. This was followed by a group discussion in which the variability in ratings was discussed in order to calibrate the ratings, so that all had an equal understanding of what constituted weak, moderate, and strong evidence, as well as how to arrive at a category score and overall case rating score. The remainder of the two-day training was spent evaluating 4-5 additional cases and conducting consensus discussions so that raters could best prepare for rating actual cases individually before engaging in a discussion with the remaining members in their group and making their final ratings.[6]

Study oversight and monitoring

Research team members were on site for the entire time during which ratings were conducted in the fall of 2012. Two members of the research team oversaw the rating teams and assigned cases for each day, while a third team member ensured materials were sufficient for scoring and assisted in checking in the data at the end of each consensus session (also checking for missing data). Depending on the complexity of the case as estimated by the researchers,

approximately two (2) to thirteen (13) cases were provided to evaluators in any given eight-hour day.

Consensus process

After half of the day's cases had been rated by all individual evaluators (evaluators were provided with 'morning' and 'afternoon' cases), a member of the research team facilitated a consensus discussion that began with raters (one at a time) providing their scores for all six categories of evidence followed by their overall case strength rating (down a column) that were transferred to a white board by the researcher. The facilitator and group reviewed the rows across, noting discrepancies of two points or more. The research protocol required that when such a discrepancy was found between any two evaluators within the team, or when the raters differed in their belief that a certain type of evidence was present or not, a facilitated discussion among evaluators was necessary. The purpose of this discussion was not to force raters to come up with the same scores. Instead, the purpose was to ensure that all raters had seen and/or considered all evidence thoroughly because of the limited time allotted to review the case (which would not necessarily be the case if the evaluators were working in their formal capacities).

The case evaluators were provided with case files stripped of case dispositions, and other necessary data, so as not to influence their determination of the case strength. All of the photo array cases involving identified suspects were assigned to two groups of raters (4 in each group) on a given day. The first group was provided with the cases inclusive of the photo array and associated pick type (but they were blind to the lineup presentation method). The second group examined the same cases, but all photo array information was redacted from the case altogether (including case details about the photo array, the photo array printout and associated pick types). Thus, their ratings were based on evidence that did not include the fact that a witness had

identified the suspect from a lineup. The results were virtually identical whether or not the photo array information was included, so we present the results averaged across that manipulation.

Results

The question of interest concerns the posterior probability of guilt (using expert ratings of evidentiary strength as a proxy) for suspects identified from simultaneous and sequential lineups in the Austin field study. As indicated earlier, lab-based ROC analyses (which usually find a simultaneous superiority effect) predict that the posterior probability of guilt – and therefore, average ratings of evidentiary strength (a proxy for "guilt") – will be higher for suspects identified from a simultaneous lineup. By contrast, using filler picks, the opposite prediction would be made (i.e., the posterior probability of guilt will be higher for suspects identified from a sequential lineup). The results again supported the prediction made by the lab-based ROC analysis. More specifically, the average evidentiary strength rating for the suspects identified from a simultaneous lineup (see Table 4) was 4.10, whereas the average rating of a suspect identified from a sequential lineup was 3.56, a difference that was statistically significant, $t(50) = 2.17$, $p = .035$, and which represents a medium effect size (Cohen's d =.61). The differences in the average ratings for filler picks and no picks from simultaneous and sequential lineups were small and did not approach significance[7]. Figure 1 summarizes the main results from Study A and Study B. Taken together, these results point to a simultaneous superiority effect in the real world AJS field data.

Discussion

The AJS field study presented a rare opportunity to evaluate the effectiveness of simultaneous and sequential lineups in the real world. In that study, actual eyewitnesses were randomly assigned to lineup type, and double-blind administration[8] was used. Moreover, overall

suspect choosing rates fortuitously turned out to be similar for both lineup types (unlike in lab studies, where suspect choosing rates are often lower for sequential lineups). That unexpected result made it possible to directly compare the diagnostic performance of the two lineup procedures while avoiding the complexities that arise when suspect choosing rates differ (in which case ROC analysis is required to meaningfully compare lineup procedures). When suspect choosing rates are the same, one need not resort to ROC analysis because the posterior odds of guilt (a close relative of the diagnosticity ratio) directly indicates which lineup procedure has a higher correct ID rate and a lower false ID rate. Using case outcomes and, separately, using expert ratings of evidentiary strength both as proxies for guilt, the AJS field data indicate that the posterior odds of guilt are higher for suspects identified from simultaneous lineups compared to sequential lineups. This result will likely be surprising to some, but it is nevertheless highly consistent with recent lab-based ROC data suggesting that sequential lineups make it harder for eyewitnesses to tell the difference between innocent and guilty suspects.

The applied implications of our findings are far reaching. It seems fair to say that the primary motivation for reforming the standard simultaneous lineup procedure has been to reduce mistaken false IDs of innocent suspects. The fact that in lab-based studies, sequential lineups typically yield a lower false ID rate (in addition to a lower correct ID rate) compared to simultaneous lineups has been interpreted to mean that the same result would likely be true in the real world. However, this does not appear to be the case. If we assume that the overall rates of choosing suspects were the same for simultaneous and sequential lineups in the AJS field study (as the data indicate), then the results reported here suggest that the sequential procedure is, if anything associated with a *higher* false ID rate in the real world. This is a sobering conclusion given that the International Association of Chiefs of Police has crafted a model policy endorsing

the sequential procedure and emphasizing that the simultaneous procedure be avoided whenever possible. Indeed, up to 30% of law enforcement agencies that use photo arrays have already switched (perhaps prematurely) to using the sequential procedure (Police Executive Research Forum, 2013), largely because sequential lineups lower the false ID rate in lab studies (and perhaps also because the filler pick rate for sequential lineups was lower in the AJS field study).

Why have years of laboratory studies found that the sequential procedure reduces the false ID rate, whereas the same result was not observed in the AJS field study? Did the lab studies get it wrong? A major difference in laboratory versus field settings has to do with fidelity or the extent to which lab studies can mimic conditions of the real-world. One criticism of lab studies for example is that the consequences associated with decision-making errors (especially choosing an innocent suspect) are much lower than in real-world settings where people's lives are at stake. For this reason alone, real eyewitnesses may be more cautious (i.e., more conservative) than participants in a lab study. In addition, the AJS field study used special instructions that were clearly designed to encourage conservative responding. For example, in addition to the standard instruction typically used in lab studies (namely, "the person who committed the crime may or may not have been included in the lineup"), the AJS field study also included instructions telling witnesses that they "did not have to make an identification" and that "the investigation would continue even if they did not identify someone." Such instructions are by no means unique to this study and are often used by law enforcement agencies. As noted by Wells et al. (2011), instructions like these "…helped make sure that the witness would not feel undue pressure to make an identification" (p. 9). That is simply another way of saying that the instructions helped to induce conservative responding. The fact that lineup instructions can be used to bring about a more conservative decision criterion has been noted by others (Clark, 2005;

Brewer et al., 2005), but the point does not appear to be widely appreciated in the eyewitness identification literature. Beyond instructions, the inclusion of a "not sure" response option in the AJS field study likely yielded even more conservative responding by siphoning off low-confidence IDs that would have otherwise occurred. The fact that deliberate steps were taken to induce conservative responding most likely explains why overall suspect choosing rates were rather low in the AJS field study (and why choosing rates did not differ for simultaneous and sequential lineups).

The fact that the overall suspect choosing rate associated with a particular lineup procedure is under the control of policymakers (and hence a “system variable”) should be emphasized because, according to one theory (Lindsay & Wells, 1985; Wells, 1984), witnesses presented with a simultaneous lineup experience pressure to make a "relative judgment." That is, they experience pressure to ID the lineup member who looks most like the perpetrator. However, as just described, pressure to make an ID can be easily reduced – or increased for that matter – by a variety of simple methods (e.g. changes in protocol such as offering an unsure option and noting that the suspect may not be in the photo array). The use of these methods will reduce suspect choosing rates for both lineup procedures and may also have the fortuitous effect of producing equivalent suspect choosing rates by effectively cancelling out any extra pressure to choose that is theoretically associated with a relative judgment strategy (thereby erasing the lower suspect choosing rate often associated with sequential lineups in lab studies). Indeed, that seems to be what happened in the AJS field study. The results of this study suggest that when standardized instructions are used to induce more conservative responding, the pressure to choose from simultaneous lineups matches that of sequential lineups. Under those conditions, simultaneous lineups appear to be diagnostically superior to sequential lineups (see Figure 1).

What would the implications of our findings be for jurisdictions in which suspect choosing rates were thought to be higher for simultaneous than sequential lineups (as is often true in lab studies)? Might sequential lineups be preferred under those conditions because of their lower false ID rates? In our view, the answer is clearly "no." A jurisdiction that uses simultaneous lineups and that wishes to reduce the false ID rate (and is willing to tolerate the loss of correct IDs that will also occur) has two choices: (1) switch to the diagnostically inferior sequential lineup procedure (which induces conservative responding while also making it harder for eyewitnesses to tell the difference between innocent and guilty suspects), or (2) stick with the simultaneous procedure and take steps to induce more conservative responding (which would reduce the false ID rate without making it more difficult for eyewitnesses to tell the difference between innocent and guilty suspects). It would only make sense to switch to the sequential procedure if the overall suspect ID rate were a fixed, immutable variable. In truth, it is to a large extent a manipulable (system) variable.[9] That being the case, there is never a reason to switch to a diagnostically inferior lineup procedure to achieve a lower false ID rate because that approach depresses the correct ID rate more than is necessary to achieve the desired outcome. A better approach would be to induce more conservative responding using the diagnostically superior procedure, which achieves the desired outcome while also maintaining the highest possible correct ID rate. More conservative responding can be achieved before the fact by using cautionary instructions, which causes witnesses to withhold low-confidence IDs that they might otherwise make, or it can be achieved after the fact by taking confidence ratings and only counting IDs made with some criterion level of confidence (such as high confidence). These two strategies are theoretically identical in that both result in the withholding of low-confidence IDs that would otherwise result in higher correct and false ID rates. Yet another complementary

approach to reducing the false ID rate without switching to a diagnostically inferior lineup procedure would be to require police investigators to provide greater justification for including a particular person as a suspect prior to proceeding with the lineup procedure (thereby reducing the chances that an innocent person would end up in a lineup in the first place).

In Phase 1 of the AJS field study (Wells et al., 2014), suspect ID rates were similar for simultaneous and sequential lineups, but filler ID rates were lower for sequential lineups (though not significantly so). As noted earlier, a filler ID does not endanger the identified individual and is therefore not treated as the equivalent of a false ID. Nevertheless, Steblay et al. (2011) argued that a filler ID from a target-absent lineup "spoils" a witness should the real perpetrator be captured and placed in a different lineup at a later time. The fact that sequential lineups are less likely to spoil witnesses in this way has been advanced as a separate argument in favor of that procedure. However, this is a debatable point because research shows that witnesses who make a filler ID when they are initially tested using a blank lineup (i.e., a lineup that contains only fillers) exhibit reduced accuracy compared to other eyewitnesses when they are tested again using a different lineup (Palmer, Brewer & Weber, 2012; Wells, 1984). Thus, an argument could be made that the simultaneous procedure is better not only because it reduces the risk to innocent suspects (as shown in Figure 1) but also because it provides useful information about witnesses whose IDs should be considered less trustworthy if they are tested again (namely, those who identified a filler on a previous test). Nevertheless, if policymakers were persuaded that it is important to reduce filler IDs in order to protect eyewitness credibility, one need not switch to the diagnostically inferior sequential lineup, which would achieve that goal while increasing the risk to innocent suspects. Instead, additional steps could be taken to induce even more conservative responding using the diagnostically superior simultaneous lineup.

What is it about simultaneous lineups that make them diagnostically superior to sequential lineups? A new theory about that issue was recently proposed by Wixted and Mickes (2014). The essence of their theory holds that a simultaneous lineup (but not a sequential lineup) provides immediate, diagnostically relevant information that an eyewitness can use to help identify a guilty suspect and to avoid misidentifying an innocent suspect. Specifically, a simultaneous lineup immediately reveals to the eyewitness that every person in the lineup shares certain facial features (e.g., every face is that of a clean-shaven white male in his mid-20s with short brown hair) – features that will also be shared by innocent and guilty suspects alike. Everyone in the lineup shares these features because those are the features that were used to select the fillers. Because these features are shared, they are *non-diagnostic* and therefore cannot be relied upon to tell the difference between innocent and guilty suspects. Instead, the shared features need to be discounted by the eyewitness in order to make an accurate ID based on other, non-shared features (e.g., shape of face, eyebrow thickness, etc.). Although simultaneous lineups draw attention to non-diagnostic (shared) features and thereby make it possible for eyewitnesses to attach less weight to them, sequential lineups do not because, in that procedure, faces are presented in isolation. Thus, when a sequential lineup is used, the witness will be more inclined to take into consideration shared features, making it harder to tell if a suspect is the perpetrator or not without other discriminable features.

In summary, our results suggest that when suspect choosing rates are similar, as they were in the AJS field study, the diagnostic accuracy of simultaneous lineups is higher than that of sequential lineups. The fact that filler choosing rates are also higher for simultaneous lineups turns out to be an irrelevant consideration (in agreement with a lab study that yielded data similar to that of the field study; see Table 2). The current results suggest that not only is the correct ID

rate higher for simultaneous versus sequential lineups, but also the false ID rate is lower, thereby balancing the concerns of justice perfectly (that innocent persons are not convicted and that guilty persons are). In light of these findings, it is hard to imagine why sequential lineups would be preferred to simultaneous lineups in practice.

Footnotes

[1]In keeping with actual practices, witnesses in the AJS field study were permitted to view the photos in the sequential lineup a second time if they requested it. In lab studies, by contrast, only one lap is typically allowed. Wells et al. (2014) analyzed the data two ways: first, by using the lap 1 results only (because this allowed them to compare the results to those found in lab studies where second laps are typically not allowed, so the lap 1 choices represent the final choices by the witness/victims in those studies); and second, by analyzing the results that accurately reflected how the sequential procedure was used in the field trial (and how it is typically used in field administration of sequential procedures, i.e. allowing a second lap on request). In the first analysis, filler ID rates were significantly higher for simultaneous compared to sequential lineups (although this analysis did not include the final decisions of the cases in which a second lap was actually requested, n = 37), but in the second analysis reflecting how the sequential procedure was actually used in the field trial, the difference in filler ID rates (specifically, 29 filler IDs out of 236 sequential lineups vs. 46 filler IDs out of 258 simultaneous lineups) was not significant (p = .09, though reported as p = .08 by Wells et al.). Only the latter (non-significant) result – the one that included the lap 2 decisions of the 16% of witnesses who requested a second viewing – is relevant to the performance of the sequential lineup in the real world. For this reason, our Phase 2 analysis included the final lap 2 decisions as well.

[2]These values were taken from Table 3 of Steblay et al. (2011) because those data came from published studies that used adults as subjects and used a full simultaneous/sequential by perpetrator-present/perpetrator-absent design. For the false alarm rates, we used the values representing "identification of designated innocent suspect."

[3]We make the assumption of equal base rates of target-present and target-absent lineups throughout (in which case the diagnosticity ratio = the posterior odds of guilt) for the sake of our illustrative examples, but none of our final conclusions depend on that assumption.

[4]Theoretically they could be endangered if district attorneys actually prosecuted known innocent fillers, but this has not to our knowledge ever been demonstrated.

[5]As suggested by our power analysis.

[6]Each group was made up of one police investigator, one prosecutor, one defense attorney and one judge.

[7]The higher average rating that was observed for suspect picks from simultaneous lineups should be balanced by a higher average rating for both filler picks and no picks from sequential lineups (because the guilty suspects who did not show up in sequential suspect picks should instead show up in the other two categories, increasing those ratings). However, that effect should be very small because there were many more filler picks and no picks in the original sample of 313 cases than suspect picks (thereby diluting the expected effect). Moreover, because only a random sample of these cases was selected for rating in Phase 2, the expected small difference in the average rating for filler picks and no picks from simultaneous and sequential lineups would have a wide confidence interval (one that would easily encompass the small and non-significant difference that was observed in favor of simultaneous lineups).

[8]Double blind administration is when not only the witness but also the lineup administrator is unaware of who the suspect is (the administrator is not the case detective) thereby eliminating the possibility that even an inadvertent cue could be sent to the witness during the photo array procedure.

[9]If the instructions were altered to say "too many guilty suspects are being released, so please make an ID even if you have only a slight hunch that you see the perpetrator in the lineup," then almost all witnesses would make an identification, whereas almost no one would make an ID if the instructions instead said "too many innocent suspects have been misidentified in recent years, so please don't make any ID unless you are 100% certain of being correct and could not possibly be making an error."

References

Amendola, K.L., & Slipka, M.G. (2009). Strength of evidence scale. Unpublished instrument. Police Foundation, Washington, DC.

Amendola, K.L., Valdovinos, M.D., Hamilton, E.E., Slipka, M.G., Sigler, M., and Kaufman, A. (2014). Photo arrays in eyewitness identification procedures: Presentation methods, influence of ID decisions on experts' evaluations of evidentiary strength, and follow-up on the AJS Eyewitness ID Field Study. Washington, DC: Police Foundation. http://www.policefoundation.org/sites/g/files/g798246/f/201403/FINAL%20EWID%20REPORT--Police%20Foundation%281%29-1_0.pdf

Brewer, N., Weber, N. & Semmler, C. (2005). Eyewitness identification. In N. Brewer & K. D. Williams (Eds.), *Psychology and law: An empirical perspective* (pp. 177-221). New York, NY: Guilford.

Carlson, C. A. & Carlson, M. A. (2014). An Evaluation of Perpetrator Distinctiveness, Weapon Presence, and Lineup Presentation using ROC Analysis. *Journal of Applied Research in Memory and Cognition, 3,* 45–53.

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 14*, 118-128.

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior, 29*, 395–424.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science,7,* 238-259.

Dobolyi, D. G. & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: a criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*, 345–357.

Durose, M.R., & Langan, P.A. (2003). *Felony Sentences in state courts, 2000*. (Report No. NCJ 198821). Retrieved from Bureau of Justice Statistics website: http://bjs.gov/content/pub/pdf/fssc00.pdf

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15,* 140-152.

Gronlund, S.D., Carlson, C.A., Neuschatz, J.S, Goodsell, C.A., Wetmore, S.A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221-228.

Gronlund, S. D., Wixted, J. T. & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23*, 3-10.

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

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

McQuiston-Surrett, D.E., Malpass, R.S., & Tredoux, C.G. (2006). Sequential vs. Simultaneous Lineups: A Review of Methods, Data, and Theory. *Psychology, Public Policy and Law, 12*, 137-169.

Mickes, L., Flowe, H. D. & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361-376.

Palmer, M. A., Brewer, N. & Weber, N. (2012). The Information Gained From Witnesses' Responses to an Initial "Blank" Lineup. *Law and Human Behavior, 36*, 439-447.

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

Police Executive Research Forum (2013). A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies. http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf

Steblay, N. M., & Phillips, J. D. (2010). The not-sure response option in sequential lineup practice. *Applied Cognitive Psychology, 25*, 768-774.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99-139.

Weber, N., & Perfect, T. J. (2012). Improving eyewitness identification accuracy by screening out those who say they don't know. *Law and Human Behavior, 36*, 28-36.

Wells, G. L. (1978). Applied eyewitness-testimony research: system variables and estimator variables. *Journal of Personality and Social Psychology, 12*, 1546-1557.

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*, 89–103.

Wells, G.L., Steblay, N.K., & Dysart, J.E. (2011). A test of the simultaneous vs. sequential lineup methods: An initial report of the AJS national eyewitness identification field studies. Des Moines, Iowa: American Judicature Society. Retrieved from: http://www.popcenter.org/library/reading/PDFs/lineupmethods.pdf

Wells, G. L., Steblay, N. K., & Dysart, J. (2012). Eyewitness identification Reforms: Are suggestiveness-induced hits and guesses true hits? *Perspectives on Psychological Science, 7*, 264-271.

Wells, G.L., Steblay, N.K., & Dysart, J.E. (2014). Double-Blind Photo-Lineups Using Actual Eyewitnesses: An Experimental Test of a Sequential versus Simultaneous Lineup Procedure. *Law and Human Behavior*.

Wixted, J. T. & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace Receiver Operating Characteristic analysis. *Perspectives on Psychological Science, 7*, 275-278.

Wixted, J. T. & Mickes, L. (2014). A signal-detection-based diagnostic feature-detection model of eyewitness identification. *Psychological Review, 121*, 262-276.

Table 1

Percentage of Witnesses who Picked a Suspect, Picked a Filler, or Rejected the Lineup when Simultaneous (SIM) or Sequential (SEQ) Lineups were Used in the AJS Field Trial (Wells et al., 2011).

	SIM	**SEQ**
Picked a Suspect	25%	27%
Picked a Filler	18%	12%
Rejected Lineup	57%	61%

Table 2

Percentage of Participants who Picked a Suspect, Picked a Filler, or Rejected the Lineup when Simultaneous (SIM) or Sequential (SEQ) Lineups were used in a Lab Study Reported by Carlson et al. (2008).

	Collapsed		Target Present		Target Absent	
	SIM	**SEQ**	**SIM**	**SEQ**	**SIM**	**SEQ**
Picked a Suspect	24%	31%	31%	41%	**16%**	**20%**
Picked a Filler	**37%**	**18%**	22%	20%	51%	16%
Rejected Lineup	40%	52%	47%	39%	33%	64%

Note: These data are from the Fair Condition of Carlson et al. (2008), which is the one condition that yielded a pattern of results similar to the AJS field study when the data were collapsed over the Target Present and Target Absent conditions.

Table 3

Number of Cases with Dispositions Provided by Research Site

Agency (Study Site)	**N**	**Guilty**	**Not Prosecuted**	**Total**
Austin, TX	143	67	76	143
San Diego, CA	24	5	19	24
Tucson, AZ	69	17	52	69
Total	236	89 (38%)	147 (62%)	236

Table 4

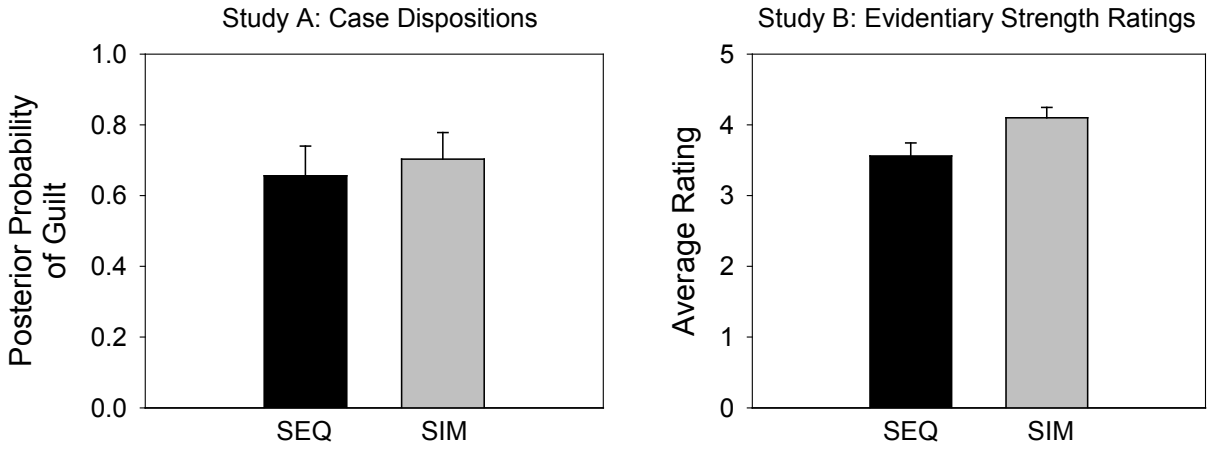
Mean Differences in Evidentiary Strength Ratings (1 – 5 scale) by Presentation Method within Pick Types across All Case Outcomes

Pick Type	Sequential	Simultaneous	t-test, significance
No pick	2.76 (29) SD 1.40	2.89 (35) SD 1.32	n.s.
Suspect	3.56 (30) SD 1.00	4.10 (22) SD 0.69	t(50) = 2.17 p = .0347
Filler	2.74 (16) SD 1.21	2.87 (19) SD 1.36	n.s.
	Total n (75)	**(76)**	

Figure Captions

Figure 1. Results of Study A (Case Dispositions) and Study B (Evidentiary Strength Ratings). The difference obtained in Study A was not statistically significant (although trended in favor of the simultaneous procedure), whereas the difference obtained in Study B was statistically significant. The results of both studies are consistent with lab-based ROC analyses suggesting that simultaneous (SIM) lineups are diagnostically superior to sequential (SEQ) lineups. Error bars represent standard errors.

Figure 1


Study A: Case Dispositions
Posterior Probability of Guilt
1.0
0.8
0.6
0.4
0.2
0.0
SEQ
SIM
Study B: Evidentiary Strength Ratings
Average Rating
5
4
3
2
1
0
SEQ
SIM