

The Role of Site Variance in the American Judicature Society Field Study Comparing Simultaneous and Sequential Lineups

Karen L. Amendola¹ · John T. Wixted²

Published online: 17 December 2015
© Springer Science+Business Media New York 2015

Abstract

Objectives Police departments often use photo lineups for eyewitness identification purposes. A widely adopted lineup reform designed to reduce eyewitness misidentifications involves switching from the standard *simultaneous* photo presentation format to a *sequential* format. These two lineup procedures were recently tested in the American Judicature Society (AJS) field study, which was conducted in four different police jurisdictions. The results from two phases of that investigation reached opposite conclusions as to which lineup procedure is superior, and the purpose of our current investigation was to elucidate the role of site variance in shaping those contrasting conclusions.

Methods In previous analyses, the field study data were either (1) aggregated across all four study sites or (2) drawn from only one study site (Austin, Texas). Here, we analyze the data separately for the Austin study site, where 69 % of the eyewitnesses were tested, and the other three study sites combined, where 31 % of the eyewitnesses were tested.

Results The results indicate significant site variance between the Austin and non-Austin study sites. In addition, the results suggest that aggregating the data across sites played a determinative role in creating the apparent disagreement about which lineup procedure is diagnostically superior.

Conclusions Once large differences across the AJS study sites are taken into consideration, there is no longer any disagreement about which lineup procedure is superior. The simultaneous procedure is diagnostically superior to the sequential procedure, but the sequential procedure sometimes induces more conservative responding (a result that can and often does masquerade as diagnostic superiority).

✉ Karen L. Amendola
kamendola@policefoundation.org

✉ John T. Wixted
jwixted@ucsd.edu

¹ Police Foundation, 1201 Connecticut Avenue, NW, Suite 200, Washington, DC 20036-2636, USA

² Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA

Keywords Eyewitness identification · Sequential lineups · Simultaneous lineups

Of the 330 wrongful convictions that have been overturned by DNA evidence since 1989, more than 70 % were attributable, at least in part, to eyewitness misidentification (Innocence Project 2015). A number of research-based recommendations have been made to reduce eyewitness identification errors, and one particularly impactful, albeit controversial, recommendation concerns the way in which the police conduct a photo lineup. A photo lineup consists of one suspect, who is either innocent or guilty, and five similar fillers, who are all known to be innocent (Fig. 1). A witness presented with a photo lineup can (1) identify a suspect (suspect ID), (2) identify a filler (filler ID), or (3) reject the lineup (no ID). To reduce misidentifications of innocent suspects, some eyewitness identification researchers have long recommended that police departments switch from using the simultaneous photo lineup, in which the photos are shown all at once, to using a sequential procedure, in which the photos are instead shown one at a time (Lindsay and Wells 1985).

The simultaneous and sequential lineup formats were recently compared using actual eyewitnesses to a crime in an investigation known as the American Judicature Society (AJS) field study (Wells et al. 2011, 2015a). The AJS field study was carried out in four different police jurisdictions and involved two phases. In Phase 1, response outcomes (suspect IDs, filler IDs and no IDs) associated with simultaneous and sequential lineups were compared (Wells et al. 2015a), whereas in Phase 2, suspects who were identified from simultaneous and sequential lineups (i.e., suspect IDs) were compared in terms of both case outcomes (adjudicated guilty vs. not prosecuted) and expert ratings of independent incriminating evidence documented in their case files (Amendola and Wixted 2015a). The results from Phase 1 were interpreted as favoring the sequential procedure, but the results from Phase 2 were interpreted as supporting the simultaneous procedure. The purpose of this article is to examine the previously unrecognized role of site variance in shaping the conclusions that were reached in the Phase 1 and Phase 2 analyses.

Fig. 1 A typical 6-pack photo lineup consists of one suspect (who may be innocent or guilty) and five similar fillers (who are all known to be innocent). A witness can identify the suspect (a suspect ID), identify a filler (a filler ID), or reject the lineup (no ID)



In an effort to gauge the diagnostic accuracy of the competing lineup procedures, eyewitness identification researchers have long relied on a statistic known as the *diagnosticity ratio* (correct ID rate/false ID rate), which indicates how likely it is that an identified suspect is guilty. Although the issue is contested (e.g., Clark 2012; Gronlund et al. 2009), some meta-analytic reviews of the mock-crime literature have concluded that the diagnosticity ratio is generally higher for sequential lineups (Stebly et al. 2001, 2011). Based on this interpretation of the empirical literature, and driven by nationwide advocacy by researchers who regard the higher diagnosticity ratio as evidence of a “sequential superiority effect”, approximately 30 % of law enforcement agencies in the US that use photo lineups have now adopted the sequential procedure (Police Executive Research Forum 2013). However, a case was recently made that the diagnosticity ratio, while being an intuitively reasonable measure to use, does not actually indicate which lineup procedure is diagnostically superior (Wixted and Mickes 2012). A National Academy of Sciences Committee on eyewitness identification endorsed this argument and further indicated that, once it is understood that the diagnosticity ratio is not the right measure to use, there is currently no compelling evidence of a sequential superiority effect in mock-crime laboratory studies (National Research Council 2014).

Despite the recent National Academy report, law enforcement agencies continue to make the switch to the sequential procedure. One reason may be that the recent revelation about the diagnosticity ratio has not been subjected to a national advocacy campaign (unlike what occurred when the evidence seemed to favor the sequential procedure), or perhaps simply not enough time has elapsed since the National Research Council (NRC) report. However, another important factor may be that the limitations of previous research are somewhat complicated to understand. The key to understanding, however, lies in grasping the difference between *response bias* and *discriminability* (Green and Swets 1966; Macmillan and Creelman 2005). The difference between these two aspects of recognition memory performance bears not only on the interpretation of the results of mock-crime lab studies but also on the interpretation the results of the AJS field study.

Response Bias Versus Discriminability

Response bias refers to the inclination of eyewitnesses to make an ID from a lineup, whereas discriminability refers to the ability of eyewitnesses to distinguish between innocent and guilty suspects. For example, if after viewing a mock crime video, one group of participants was instructed to make an ID from the lineup even if they had to guess, whereas another group was instructed to make an ID only if they were certain of being correct, the first group would exhibit a more liberal response bias (i.e., they would make more IDs, leading to a higher correct ID rate and a higher false ID rate) than the second, more conservative, group. Despite that difference in performance, the two groups as a whole would not differ in their ability to discriminate innocent from guilty suspects. By contrast, if one group was allowed to view the mock-crime video five times, whereas the other group was allowed to view the video only once, then the first group would be expected to exhibit higher discriminability (i.e., a greater ability to distinguish between innocent and guilty suspects) than the second. The first group might, for example, exhibit both a higher correct ID rate and a lower false ID rate than the second.

Switching from one lineup procedure (e.g., simultaneous) to another (e.g., sequential) can influence response bias, discriminability or both. An effect on response bias would be evident if both the correct ID rate and false ID rate were lower for one procedure than the other, as is often true of sequential lineups compared to simultaneous lineups. Thus, it is widely agreed that switching to the sequential procedure often has at least one effect, namely, it induces more conservative responding (Stebly et al. 2011).

An important consideration that has only recently come to be understood in the field of eyewitness identification is that a natural consequence of more conservative responding (in addition to the fact that the correct and false ID rates decrease) is that *the diagnosticity ratio increases* (Rotello et al. 2015; Gronlund et al. 2014; Wixted and Mickes 2012, 2014). Critically, this occurs whether more conservative responding is induced for the simultaneous procedure (e.g., using instructions that encourage eyewitnesses not to make an ID unless they are confident of being correct) or more conservative responding is induced by switching to the sequential procedure. The diagnosticity ratio continues to increase as responding becomes ever more conservative, all the way to the point where both the correct and false ID rates approach 0, in which case administering a lineup would be practically useless even though the diagnosticity ratio would be very high (Wixted and Mickes 2014). Thus, achieving the highest possible diagnosticity ratio by inducing ever more conservative responding is not a goal that is worth pursuing.

The eyewitness ID procedure that should be advocated by researchers is the one that best enables eyewitnesses to (1) identify a guilty suspect in a lineup (when presented with a target-present lineup) and (2) avoid identifying an innocent suspect in a lineup (when presented with a target-absent lineup). In other words, the procedure that should be recommended is the one that yields the highest discriminability. The diagnosticity ratio does not indicate which procedure yields higher discriminability, so another approach is needed.

Receiver Operating Characteristic Analysis

A well-established methodology for determining which of two diagnostic procedures yields the highest discriminability is called receiver operating characteristic (ROC) analysis (Green and Swets 1966; Macmillan and Creelman 2005). As noted by the National Academy committee on eyewitness identification, “ROC analysis represents an improvement over a single diagnosticity ratio” (National Research Council, p. 80). An ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (the correct ID rate in the case of lineups) against the false positive rate (the false ID rate) at various threshold settings as a measure of sensitivity (as illustrated in Fig. 3).

ROC analysis has its roots in World War II radar and electrical engineering, and it has since been widely adopted in fields such as psychology, medicine, and biometrics, to name a few. Although still a new development in the field of eyewitness identification, recent ROC analyses in mock crime laboratory studies have often found that the simultaneous procedure yields higher discriminability than the sequential procedure (Carlson and Carlson 2014; Dobolyi and Dodson 2013; Gronlund et al. 2012; Mickes et al. 2012). These findings suggest that the simultaneous procedure is diagnostically superior to the sequential procedure, which is the opposite of the conclusion that has been reached based on the diagnosticity ratio. With these essential background considerations in mind, we return now to a consideration of the role of site variance in the AJS field study.

Site Variance in the AJS Field Study

Regardless of what lab studies suggest, the more important question concerns the performance of simultaneous and sequential lineups in the real world. Which lineup procedure is diagnostically superior when actual eyewitnesses to a crime are tested? As noted earlier,

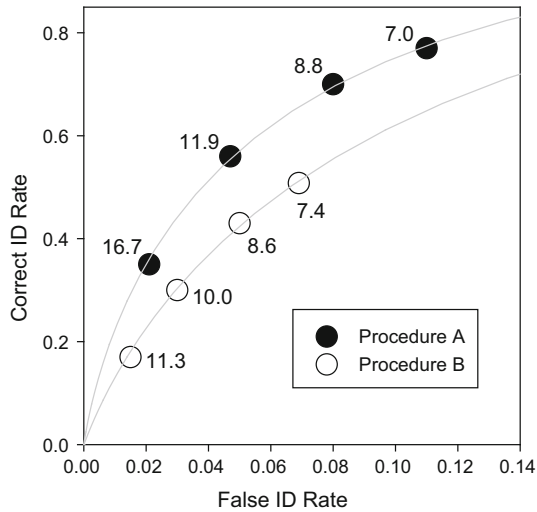


Fig. 3 Illustration of receiver operating characteristic plots for two hypothetical lineup procedures. Each lineup procedure is constrained to yield correct and false ID rates that fall on a curve as responding changes from being very conservative (lower leftmost point of each procedure) to being very liberal (upper rightmost point for each procedure). Values shown next to each data point indicate the diagnosticity ratio (correct ID rate/false ID rate) for that point. In this example, Procedure A is diagnostically superior to Procedure B because for any given false ID rate, Procedure A can achieve a higher correct ID rate. If only a single ROC point is computed for each procedure and are then compared using the diagnosticity ratio (as was done in the vast majority of mock-crime lab studies comparing simultaneous and sequential lineups), the diagnostically inferior lineup procedure could be misconstrued as being the superior procedure (e.g., imagine computing only the rightmost ROC point for each procedure and comparing them using the diagnosticity ratio)

this question was addressed by the AJS field study (Wells et al. 2011, 2015a), which was carried out in four different police jurisdictions: (1) Charlotte-Mecklenburg County, North Carolina; (2) Tucson, Arizona; (3) San Diego, California; and (4) Austin (Travis County), Texas. In this field study, 494 actual eyewitnesses to a crime were randomly assigned to either a simultaneous lineup or a sequential lineup.

In Phase 1 of the study, the data were analyzed in terms of the number of suspect IDs, filler IDs, and no IDs that were made using simultaneous and sequential lineups (Wells et al. 2015a). We have argued that this is not a sufficient means for determining diagnostic superiority (Amendola and Wixted 2015a) because, to convincingly address that issue, one must also know the “ground truth” of which identified suspects are actually guilty (or at least are more likely to be guilty) when identified from one lineup procedure compared to the other. However, in field studies (unlike in mock-crime studies), it is not immediately known if suspect IDs involve innocent or guilty suspects. In an effort to get around that problem, the main focus in Phase 1 of the AJS field study was on filler IDs. Because fillers are known to be innocent, Wells et al. (2015a) made the assumption that if one lineup procedure results in more IDs of innocent fillers, it probably results in more IDs of innocent suspects as well. Although we have previously argued that this assumption may not be accurate (see Amendola and Wixted 2015a), whatever one’s position on that issue might be, it is important to determine whether site variance affected the conclusions that Wells et al. (2015a) drew based on the filler ID rates.

Phase 2 of the AJS field study focused on suspect IDs.¹ In this phase of the investigation, an effort was made to determine if the identified suspects were innocent or guilty (i.e., an effort was made to get at the “ground truth” of guilt or innocence). This was accomplished in two ways: first by examining *case outcomes* (i.e., by asking whether the identified suspects were ultimately adjudicated guilty or not) and, second, by obtaining *expert ratings* of the strength of independent incriminating evidence documented in the case files of the identified suspects (Amendola, et al. 2014; Amendola and Wixted 2015a). Because suspect IDs—especially innocent suspect IDs—are far more consequential than filler IDs,² this approach more directly addresses the question of whether simultaneous or sequential lineups result in fewer false IDs of the innocent and more correct IDs of the guilty. Amendola and Wixted (2015a) collected ratings of evidentiary strength from those who evaluate these cases daily, i.e., police investigators, prosecutors, defense attorneys, and judges. In our view, the evaluation of a number of categories of evidence was a better proxy for ground truth than actual case outcomes, not only because these experts examined all evidence in the case,³ but also because these experts rated the evidence in the cases independent of procedural influences that may be present in court (e.g., inadmissibility of evidence, failure of witnesses to appear, etc.). While we recognize that using evidentiary strength ratings as a proxy for guilt is not without its limitations, we believe it comes significantly closer to estimating ground truth than do actual case outcomes, which are also mostly reached through plea agreements.

How was the issue of site variance addressed in the Phase 1 and Phase 2 investigations? In their Phase 1 investigation, Wells et al. (2015a) briefly addressed the issue using a statistical test and concluded that there was no evidence of site variance (as described in more detail below). For that reason, their analyses were based on data that were aggregated across all four study sites. By contrast, when designing their Phase 2 investigation, Amendola et al. (2014) were quite concerned about site variance because it did not appear that the methodological or analytical approach used by Wells et al. (2011) included controls for site variation, such as relying on a randomized block experiment, as has been recommended for multi-site trials (see Weisburd and Taxman 2000). Because of that concern, and because the majority of the data generated in Phase I were from Austin (340 out of 494 lineups were administered there), Amendola et al. (2014) limited their Phase 2 expert ratings study to that site alone.^{4,5} Had the Phase I experiment been designed to

¹ In actuality there were two purposes of the Phase II study; the first was to examine the case outcomes (both actual judicial outcomes of the cases, and rated evidentiary strength), and the latter was to experimentally test whether the identification of a suspect in a lineup influenced criminal justice decision makers in their ratings of the other evidence in the case.

² Because known innocents are used as fillers in lineups, they are not at risk of prosecution.

³ Teams of case evaluators made up each of one police investigator, prosecutor/DA, defense attorney, and judge rated the same case independently and then discussed their reasons for these ratings. After that, all raters were allowed to change their ratings before we computed an average score for evidentiary strength. Additionally, the teams of raters stayed the same in a given day in which from 3 to 12 cases were rated on average, but every day, teams were changed based on the availability of the raters, so that the teams were well counterbalanced.

⁴ Amendola et al. (2014) decided to conduct their experimental study strictly in Austin because (a) the vast majority of cases were drawn from Austin; (b) the limited number of cases in two of the sites; (c) cost-benefit analysis of conducting the study in sites where there were insufficient cases for independent analysis; and (d) minimization of error that could be induced from site variance.

⁵ One specific concern was the variance in completion rates for lineups across sites, with a much more rapid rate of study completion in Austin as compared to other sites, indicating some hesitancy in fully complying with the experiment or its protocols in other sites (e.g. in one site there were technical concerns over

control for site variance (e.g., using a randomized blocked design and analysis approach), the decision to limit the study to one site would not have been considered as important. We turn now to a detailed consideration of the role of site variance in Phase 1 and Phase 2 of the AJS field study.

Phase 1 Results

Table 1A, B show the Phase 1 results aggregated across the 4 study sites, which is how the data were analyzed by Wells et al. (2015a). Table 1A shows the raw frequency counts, and Table 1B shows the same data expressed as proportions. A Chi square test of independence performed on the frequency counts was far from significant, $\chi^2(2) = 2.94, p = .230$. Thus, the two lineup procedures did not differ in terms of the relative frequencies of the three possible eyewitness decisions. However, Wells et al. (2015a) performed separate tests on each of the three simultaneous-versus-sequential pairwise proportions (one for suspect IDs, one for filler IDs, and one for no IDs), and the p values for these tests are shown in the rightmost column of Table 1B. As indicated by the arrow, for one of those tests, there is a non-significant ($p = .09$) trend suggesting a higher proportion of filler IDs being made for simultaneous lineups (.18) compared to sequential lineups (.12).⁶ Although not significant, this empirical trend has been featured in continued advocacy for the sequential procedure based on the assumption that, if the effect is real, it means that sequential lineups reduce the risk that an innocent suspect will be misidentified.

What do the data suggest when they are *not* aggregated across study sites? The subset of 340 cases from the Austin study site (69 % of the AJS field data) was provided to us by Wells and colleagues for our Phase 2 investigation, which allowed us to directly compute the response outcome data from the collective set of non-Austin study sites by simple subtraction. Having that information in hand made it possible to ask whether there is evidence of site variance between the Austin site (340 cases) and the non-Austin sites (collectively, 154 cases). One might also wonder if there is evidence of site variance among the three non-Austin study sites, but (1) we do not have the data broken down separately for those sites; (2) any effect of site variance among those three sites would be hard to detect given that the number of observations from each site would be small; and (3) the question of site variance among the three non-Austin sites is independent of our main focus, which concerns the possibility of site variance between the Austin site (where our expert ratings study was conducted) versus the three non-Austin sites.

To investigate the possibility of site variance, we first analyzed the data in a manner similar to the site-variance test performed by Wells et al. (2015a). As noted above, they collapsed across lineup format (simultaneous vs. sequential) and examined the distribution of suspect IDs, filler IDs, and no IDs as a function of study site. Because no significant differences were observed, they aggregated the data across study sites for all of their

Footnote 5 continued

software issues thereby reducing willingness to rely on the software for administering lineups, as well as slowing down the rate of completed pristine lineups).

⁶ After viewing the six photos in a sequential lineup, a witness may request a second lap through the photos before making a final identification decision. A second lap is typically not allowed in lab studies, but it is almost invariably allowed in actual practice. About 16 % of witnesses in the AJS field study requested a second lap, and all of our analyses include their final decisions. Wells et al. (2015a) reported that the filler ID difference between simultaneous and sequential lineups was significant when the analysis was limited to lap 1 decisions, but that result is relevant only to lab studies, not to how sequential lineups are used in practice and were actually applied in these cases in the field.

Table 1 Identification (ID) frequencies (A), proportions (B), and Chi Square analysis for Simultaneous (SIM) and Sequential (SEQ) Lineups in the AJS Field Study aggregated across the 4 study sites

<i>N</i> = 494			Proportions			
Pick type	SEQ	SIM	Pick type	SEQ	SIM	<i>p</i>
(A)			(B)			
Suspect ID	65	67	Suspect ID	.28	.26	.67
Filler ID	29	46	Filler ID	.12	.18	.09 ←
No ID	142	145	No ID	.60	.56	.39

$\chi^2(2) = 2.94, p = .230$

Note the arrow designates the finding that has been interpreted to indicate a sequential advantage

Table 2 Identification (ID) frequencies (A), proportions (B), and Chi Square analysis for Austin versus non-Austin sites in the AJS Field Study (aggregated across simultaneous and sequential lineups)

Pick type	Austin	Non-austin	Pick type	Austin	Non-Austin
(A)			(B)		
Suspect ID	93	39	Suspect ID	.27	.25
Filler ID	47	28	Filler ID	.14	.18
No ID	200	87	No ID	.59	.56

$\chi^2(2) = 1.58, p = .452$

subsequent analyses comparing simultaneous and sequential lineups. Using the same approach to testing for evidence of site variance in Table 2 here (i.e., comparing the Austin vs. non-Austin site data collapsed across lineup format), we also find no hint that it played any role, $\chi^2(2) = 1.58, p = .452$. However, the purpose of the AJS field study was to compare simultaneous and sequential lineups. Thus, with regard to the possibility of site variance, the question of interest is whether the distribution of suspect IDs, filler IDs, and no IDs differed for simultaneous and sequential lineups as a function of study site. The only way to answer that question is to examine the site-specific data separately for the two lineup formats.

Table 3A, B show the relevant frequency counts separately for the Austin and non-Austin study sites, respectively (i.e., they show the same data that are aggregated across study sites in Table 1 and aggregated across lineup format in Table 2). Separate 2 × 3 Chi square tests performed on the frequency counts from the Austin site (Table 3A) and non-Austin sites (Table 3B) show that the data from Austin alone (Table 3A) do not yield any hint of a significant effect, but the data from the non-Austin sites (Table 3B), despite being based on many fewer eyewitness decisions, is significant, $\chi^2(2) = 7.88, p = .019$. These results provide clear evidence of site variance.

Do the patterns of data in Table 3A, B differ significantly from each other? The most direct way to test whether or not they do is to perform a 2 × 6 Chi square test of independence, with site (Austin vs. non-Austin) as one factor and response outcome (SIM suspect ID, SIM filler ID, SIM no ID, SEQ suspect ID, SEQ filler ID, SEQ no ID) as the other factor. The result of this test was significant, $\chi^2(5) = 11.11 p = .049$, indicating that the patterns do in fact differ significantly from each other.

Table 3C, D (the corresponding proportions) clarify what the difference between the two study sites is. In Austin, the proportions of suspect IDs, filler IDs and no IDs are

Table 3 Identification (ID) frequencies (A, B), proportions (C, D), and Chi Square analysis for Simultaneous (SIM) and Sequential (SEQ) Lineups in the AJS Field Study separately for the Austin study site (A, C) and the other three non-Austin study sites combined (B, D)

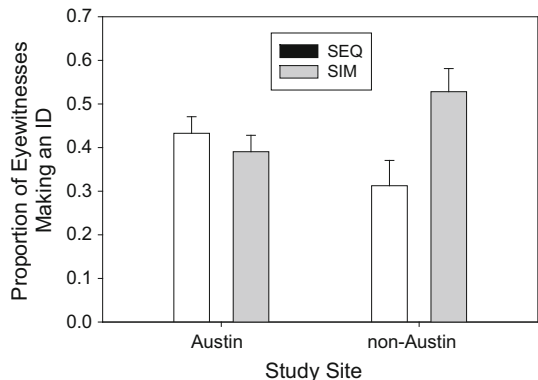
Austin (N = 340)				Non-Austin (N = 154)			
Pick Type	SEQ	SIM		Pick Type	SEQ	SIM	
(A)				(B)			
Suspect ID	52	41		Suspect ID	13	26	
Filler ID	22	25		Filler ID	7	21	
No ID	97	103		No ID	45	42	
$\chi^2(2) = 1.66, p = .436$				$\chi^2(2) = 7.88, p = .019^*$			
Pick Type	SEQ	SIM	p	Pick Type	SEQ	SIM	p
(C)				(D)			
Suspect ID	.30	.24	.20	Suspect ID	.20	.29	.22
Filler ID	.13	.15	.61	Filler ID	.11	.24	.046*
No ID	.57	.61	.43	No ID	.69	.47	.008**

* $p < .05$; ** $p < .01$

similar for simultaneous and sequential lineups. In the non-Austin sites, by contrast, it is evident that witnesses were far more reluctant to make an ID from sequential lineups compared to simultaneous lineups. This is most easily appreciated by considering the fact that, in the non-Austin sites, 69 % of the eyewitnesses tested using the sequential procedure made no ID (neither a suspect ID nor a filler ID), whereas only 47 % of the eyewitnesses tested using the simultaneous procedure made no ID. In other words, responding was significantly more conservative for sequential lineups in the non-Austin sites, whereas response bias was similar for simultaneous and sequential lineups in the Austin site. We do not know why the sites differed in this respect, but it is clear from the data that they did. Similar variability has often been observed across lab studies (i.e., in lab studies, sequential lineups often, but not always, yield more conservative responding).

The difference in response bias across study sites can be most clearly illustrated if the two ID types (suspect IDs and filler IDs) are aggregated together. As shown in Fig. 4, there is no apparent difference in response bias for simultaneous versus sequential lineups in

Fig. 4 Proportion of eyewitnesses who made an ID (suspect ID or filler ID) from simultaneous (SIM) and sequential (SEQ) lineups administered at the Austin and non-Austin study sites. Error bars represent the standard error of a proportion



Austin (43 % of eyewitnesses tested using a sequential lineup made an ID of some kind, and 39 % of eyewitnesses tested using a simultaneous lineup did so). In the *non*-Austin sites, by contrast, there is a large difference in response bias, with sequential lineups clearly yielding a more conservative response bias than simultaneous lineups (only 31 % of eyewitnesses tested using a sequential lineup made an ID, whereas 53 % of eyewitnesses tested using a simultaneous lineup did so). The *non*-Austin eyewitnesses were more significantly liberal than their Austin counterparts when tested using the simultaneous procedure, $\chi^2(1) = 4.48, p = .034$, and they were marginally more conservative than their Austin counterparts when tested using the sequential procedure, $\chi^2(1) = 3.07, p = .080$. The two effects combined resulted in the substantial difference in response bias between simultaneous and sequential lineups for the *non*-Austin sites.

What does all this mean for the interpretation of the Phase 1 results of the AJS field study based on the aggregated data shown earlier in Table 1A, B? The existence of site variance means that it is inappropriate to aggregate the data across study sites, as Wells et al. (2015a) did. In terms of suspect IDs, filler IDs and no IDs, the results of the AJS field study should instead be interpreted using Table 3A–D. The only conclusion that can be drawn from those data is that in the *non*-Austin sites, but not in the Austin site, responding was more conservative for the sequential procedure (a result that says nothing about which procedure is diagnostically superior).

Recall that Wells et al. (2015a) interpreted the effect of lineup format on filler IDs (non-significantly lower for the sequential procedure in the data aggregated across study sites, denoted by the arrow in Table 1B) to indicate a sequential advantage. However, it is apparent that this non-significant trend is almost entirely the result of the more conservative responding that occurred in the *non*-Austin sites. This is most easily appreciated by examining the filler ID proportions from Table 3C (for Austin), where the values for SEQ (0.13) and SIM (0.15) are nearly identical versus Table 3D (for the *non*-Austin sites), where the values for SEQ (0.11) and SIM (0.24) differ to a large degree. As indicated earlier, more conservative responding (which can be easily achieved using the simultaneous procedure, if so desired) is not an indication of diagnostic superiority. To interpret this result as being indicative of diagnostic superiority is to make the same mistake that was made with respect to the interpretation of data from lab studies, where the more conservative responding and the attendant increase in the diagnosticity ratio often induced by the sequential procedure was, for many years, interpreted to indicate a “sequential superiority effect” (Stebly et al. 2001, 2011). Although somewhat counterintuitive, more conservative responding is not an indication of diagnostic superiority. To quote from the recent National Academy of Sciences’ NRC report on eyewitness identification evidence:

Intuitively, if sequential lineups yield a higher likelihood that an identified person is guilty (as quantified by a higher diagnosticity ratio), then it seems as if that procedure yields objectively better performance. The problem with this intuition is that it fails to take into account the second of the two parameters of recognition memory models—the response bias or degree of evidence that the observer finds acceptable to make an identification. This parameter, which is distinct from discriminability, reflects the witness’ tendency to pick or not to pick someone from the lineup...Differences in pick frequency can, and generally do, lead to differences in the ratio of hit rates to false alarm rates; all else being equal, the diagnosticity ratio will be higher for a conservative bias than for a liberal bias. In other words, simply by inducing a witness to adopt a more conservative bias, it is possible to increase the likelihood that an identified person is actually guilty. Importantly, this may be true

even if the procedure yields no better, or potentially worse, discriminability (NRC, p. 55–56).

The key point is that, whether in the lab or in the field, more conservative responding is not an indication of diagnostic superiority, yet more conservative responding in the non-Austin study sites accounted for the non-significant filler ID rate trend evident in the aggregated data reported by Wells et al. (2015a). Thus, what Wells et al. (2015a) interpreted as a sequential advantage is instead an indication of more conservative responding (in the non-Austin study sites), which is precisely the same error that resulted in original notion (based on lab studies) that sequential lineups are diagnostically superior to simultaneous lineups. The superior lineup procedure is the one that yields higher discriminability (i.e., the ability to distinguish between *suspects* who are innocent or guilty), as also pointed out by the NAS Committee: “The committee concludes that there should be no debate about the value of greater discriminability—to promote a lineup procedure that brings less discriminability would be akin to advocating that the lineup be performed in dim instead of bright light” (NRC, p. 80).

Phase 2 Results

Phase 2 of the AJS field study focused on suspects who were identified from simultaneous and sequential lineups because, whether innocent or guilty, these are the individuals who are placed at risk of prosecution and, ultimately, conviction (whereas fillers who are identified are not placed at similar risk, as the prosecutors and police know these individuals to not be associated with the crime). Thus, it would be useful to know which lineup procedure does a better job of maximizing IDs of guilty suspects while minimizing IDs of innocent suspects. That is, it would be useful to know which procedure makes it easier for eyewitnesses to *discriminate* innocent from guilty suspects.

To determine whether the identified suspects in the AJS field study were innocent or guilty, the Phase 2 analysis used two approaches. First, *expert ratings* of evidentiary strength were made by law enforcement and legal professionals who examined the suspect case files, and second, *case outcomes* (i.e., ultimately adjudicated guilty or not) were examined. As noted earlier, for the evidentiary strength ratings analysis, case files were examined by a team of judges, prosecutors, defense attorneys and police detectives who rated each suspect’s file using a 1-to-5 (low-to-high) scale for independent corroborating evidence that the suspect committed the crime. The expert raters were blind to the type of lineup that was used to make the identification (see Amendola and Wixted 2015a, for details). For the case-outcome analysis, the question was whether a higher proportion of suspects identified from one lineup procedure was ultimately adjudicated guilty (by jury verdict or plea bargain) compared to the other lineup procedure.

The ratings study was conducted in Austin in coordination with the Travis County DA’s office. Thus, that analysis was inherently limited to eyewitness identifications made at that one study site, whereas the case-outcome analysis reported by Amendola and Wixted (2015a) included suspects identified across multiple study sites. Given the evidence of site variance discussed above, it would have made more sense to separate the case-outcome analysis for the Austin and non-Austin sites. We did not do that previously (as we should have, given evidence of site variance), but we do so here for the first time.

Critically, both of these measures (i.e., evidentiary strength ratings and case outcomes) are conceptually identical to the diagnosticity ratio that is routinely measured in lab studies. In each case, what is being measured are the odds that an identified suspect is

actually guilty. Thus, the issue that bedevils lab studies (namely, that more conservative responding inflates the diagnosticity ratio) is potentially problematic for these field study measures as well. More specifically, in a field study, if responding is more conservative for one lineup procedure than the other, then, all else being equal, the odds of guilt associated with suspects who are identified using the more conservative lineup procedure would be expected to exceed the odds of guilt associated with suspects who are identified using the less conservative lineup procedure (for the same reason that the diagnosticity ratio in lab studies is expected to be higher for the lineup procedure that induces more conservative responding). Such a result favoring the more conservative procedure would not be an indication of lineup superiority. Instead, it might simply reflect the fact that responding was more conservative using that procedure. An “odds of guilt” measure directly indicates which lineup procedure yields higher discriminability *only when response bias does not differ between the two procedures* (Amendola and Wixted 2015a). Response bias did not differ in Austin, so an “odds of guilt” measure—such as case outcomes or evidentiary strength ratings—can be used to determine the diagnostically superior lineup procedure. However, response bias did differ for the non-Austin study sites, so the natural expectation is that an “odds of guilt” measure would favor the more conservative procedure (without indicating that the more conservative procedure is diagnostically superior). Although we do not know why response bias differed across study sites, given that it did, data from the different sites cannot be legitimately combined and must instead be analyzed separately.⁷

Evidentiary Strength Ratings

As indicated above, Table 3A, C show that response bias did not differ appreciably for simultaneous and sequential lineups in the Austin study site. Thus, the conditions for using an “odds of guilt” measure to identify the superior lineup procedure were in place. Our ratings study was already limited to the Austin site, so we begin with a brief consideration of those straightforward results (which are obviously not vulnerable to the issue of site variance) and then, for the first time, consider the case-outcome data separately for the Austin and non-Austin sites.

Which group of identified suspects from the Austin study site—those who were identified from simultaneous lineups or those who were identified from sequential lineups—was more likely to be associated with independent incriminating evidence according to the expert raters? That is the question we set out to address in the expert ratings study. As shown in Table 3A, 52 suspects were identified from sequential lineups and 41 from simultaneous lineups (93 in all) in Austin. However, not all of these suspects could be included in the ratings study. Based on factors beyond our control, 19 of the suspects identified from simultaneous lineups and 22 of the suspects identified from sequential lineups could not be included. For example, Texas state law required us to exclude cases involving juveniles, the DA’s office required us to exclude cases involving sexual assault, some cases were found to have irreconcilable inconsistencies in case details, and some made no mention of the suspect at all, etc. (see Amendola et al. 2014). In the end, we were left with 30 suspects identified from sequential lineups (52 initially identified—22 necessarily excluded = 30 identified sequential suspects) and 22 from simultaneous lineups (41 initially identified—19 necessarily excluded = 22 identified simultaneous suspects). To keep the numbers as high as possible, every identified suspect who could be included in

⁷ Another anticipated concern, as the limited number of cases in two of the sites would also limit statistical power.

the ratings study was included. Because eyewitnesses were randomly assigned to lineup format in the AJS field study, it seems reasonable to assume that they were randomly distributed across the various exclusionary factors as well.

With regard to the evidentiary strength ratings, the results were entirely straightforward: on average, the evidentiary strength ratings were significantly higher for the 22 simultaneous suspects (mean = 4.10, SD = 0.69) than for the 30 sequential suspects (mean = 3.56, SD = 1.00), $t(50) = 2.17$, $p = .035$. This difference corresponds to a Cohen's d of .61 (a medium effect size). Given that response bias did not differ for simultaneous and sequential lineups, this result indicates that simultaneous lineups better enable eyewitnesses to distinguish between innocent and guilty suspects (i.e., to better identify guilty suspects from target-present lineups and to avoid misidentifying innocent suspects from target-absent lineups). That is, to the extent that the ratings are indicative of actual guilt, the results indicate that simultaneous lineups yield higher discriminability than sequential lineups. This is the same conclusion reached by recent lab-based ROC analyses that were performed once it became clear that the diagnosticity ratio does not identify the superior procedure (e.g., Mickes et al. 2012). That is, both sets of results point to a simultaneous superiority effect.

Case Outcomes

Do the case outcomes analyses (adjudicated guilty vs. not prosecuted) agree with the ratings analysis? Much confusion and debate surrounds this issue, so we consider it in some detail here. After the ratings study was completed, we sought case outcome information for all 52 of the Austin suspects who were included in the ratings study. At the time, case outcome information was available for 47 of these suspects (including all 22 simultaneous suspects and 25 of the 30 sequential suspects) and was unavailable for 5 sequential suspects. All five of these sequential suspects were likely “not prosecuted” (because the DA’s office had no information about their outcomes), but this was not definite, so these cases were not initially included in the analysis reported by Amendola and Wixted (2015a). Additionally, case outcome information was provided to us for another 22 suspects who were identified from lineups administered at the other (non-Austin) study sites. Thus, definite case outcome information was available for 47 (Austin) + 22 (non-Austin) = 69 suspects in all (32 sequential and 37 simultaneous). Being unaware of the site variance reported here, Amendola and Wixted (2015a) originally reported case outcome information for these 69 suspects aggregated across study sites, and the results showed that 65.6 % of sequential suspects and 70.3 % of simultaneous suspects were adjudicated guilty (Table 4A). Thus, this analysis shows no hint of a sequential advantage, but it provides no compelling evidence for a simultaneous advantage either. However, we subsequently realized that it makes no more sense to compute this statistic aggregated across study sites than it does to compute any of the other statistics from the AJS field study aggregated across study sites.

Based on the possibility of site variance, Amendola and Wixted (2015b) subsequently reported case-outcome data for the suspects identified in Austin only (the same group that was used for our evidentiary strength ratings study). Table 4B shows the case outcome data for the 47 Austin suspects who were included in the aggregated data shown in Table 4A. The slight advantage for the simultaneous procedure over the sequential procedure evident in the aggregated data (.703 vs. .656, Table 4A) becomes noticeably larger when the data are limited to the Austin data (.773 vs. .640, Table 4B), though the difference is still not significant. The disparity grows even larger when the five sequential suspects with non-

Table 4 Case outcome frequencies for suspects identified from SEQ and SIM lineups across the four AJS study sites (A), in Austin alone but not including five sequential suspects with initially uncertain case outcomes (B), in Austin alone but now including five sequential suspects after a further inquiry into their case outcomes (C), and in the non-Austin study sites (D)

Aggregated across study sites (<i>N</i> = 69 identified suspects)			Austin (<i>N</i> = 47 identified suspects)		
Outcome	SEQ	SIM	Outcome	SEQ	SIM
(A)			(B)		
Adjudicated guilty	21	26	Adjudicated guilty	16	17
Not prosecuted	11	11	Not prosecuted	9	5
Proportion guilty	0.656	0.703	Proportion guilty	0.640	0.773
Austin (<i>N</i> = 52 identified suspects)			Non-Austin (<i>N</i> = 22 identified suspects)		
(C)			(D)		
Adjudicated guilty	16	17	Adjudicated guilty	5	9
Not prosecuted	14	5	Not prosecuted	2	6
Proportion guilty	0.533	0.773	Proportion guilty	0.714	0.600

definite case outcomes are now included in the analysis (.773 vs. .533, Table 4C). These are the data that Amendola and Wixted (2015b) reported after again making inquiries into the case outcomes of the five sequential suspects (in December of 2014) and being informed by the Travis County DA’s office that they still had no indication that any of these suspects had ever been prosecuted. It therefore seems reasonable to suppose that they never will be prosecuted, in which case all five belong in the “not prosecuted” category. For the 47 + 5 = 52 suspects identified in Austin, 53.3 % of suspects identified from sequential lineups were adjudicated guilty and 77.3 % of suspects identified from simultaneous lineups were adjudicated guilty (Table 4C), a difference that is marginally significant ($p = .077$). The effect-size for this difference in proportion guilty (.773 vs. .533) using Cohen’s *h* is .511, which corresponds to a medium effect size. Thus, for the Austin data considered separately, the ratings data show a significant result favoring the simultaneous lineup and the case-outcome data show a non-significant trend favoring the simultaneous lineup. It seems fair to conclude from this overall pattern of results that if any difference in diagnostic accuracy exists between the two lineup formats, it likely favors the simultaneous procedure.

The Debate over the Findings Reported by Amendola and Wixted (2015a)

The large increase in the apparent advantage for simultaneous lineups between the data aggregated over study sites shown in Table 4A (70.3 vs. 65.6 %) and the data from Austin shown in Table 4C (77.3 vs. 53.3 %) struck Steblay et al. (2015) as indicating that the Austin ratings data must have been biased against the sequential procedure. After all, in the absence of a selection bias, why would the case outcome data from the Austin sample (Table 4C) differ so noticeably from the case outcome data in the aggregated data set (Table 4A)? According to Steblay et al. (2015): “The evidence that the set of 52 lineups used by Amendola and Wixted is unrepresentative of the larger data set is indisputable” (p. 296). Indeed, in their Table 1, they provide expected frequencies for a random sample of

52 suspects drawn from the larger set of 69 suspects aggregated across study sites,⁸ and those expected frequencies differ markedly (and significantly) from the observed frequencies, shown here in Table 4C. However, due to site variance (a factor not considered by Steblay et al. 2015), there is no reason to expect the Austin data set to be representative of the larger data set aggregated across study sites. Moreover, the pattern observed for the Austin data set deviates from the pattern observed for the larger data set in just the way one would expect given that (1) response bias (i.e., the tendency to make an ID) is about the same for simultaneous and sequential lineups in the Austin data and (2) is much more conservative for sequential lineups than simultaneous lineups in the non-Austin data.

As noted above, the Austin case-outcome data in Table 4C show a marginally significant trend favoring the simultaneous procedure ($p = .077$). Because response bias is the same for simultaneous and sequential lineups, these results can be interpreted to mean that there is a marginally significant trend towards higher discriminability for simultaneous lineups. By contrast, the fact that more conservative responding is evident for sequential lineups in the non-Austin sites leads to the prediction that the non-Austin case-outcome data should exhibit an effect in the *opposite direction*. Again, this prediction arises because a conservative response bias inflates any odds of guilt measure. Table 4D shows the case-outcome data for the 22 non-Austin suspects who were identified from simultaneous and sequential lineups. Although the numbers are too small yield significant effects, those data do indeed show a trend in the opposite direction, now appearing to favor the *sequential* procedure. Specifically, 71.4 % of suspects identified from sequential lineups were adjudicated guilty, whereas 60.0 % of suspects identified from simultaneous lineups were adjudicated guilty. Steblay et al. (2015) were troubled by the difference between the aggregated data (Table 4A) versus the Austin data (Table 4C), but their key point is that the apparent difference between the two sets of data is entirely in line with what would be expected given the evidence of site variance in the AJS field data. In other words, the fact that the Austin data differ from the aggregated data does not reflect a mysterious bias against the sequential procedure, as Wells et al. (2015b) and Steblay et al. (2015) assumed must be the case; instead, it reflects the heretofore unrecognized existence of site variance in the AJS field study.

Discussion

Our main goal is to elucidate the previously unappreciated role of site variance in affecting the conclusions that were reached in the AJS field study about whether simultaneous or sequential lineups are diagnostically superior. Because the Austin and non-Austin study sites yielded data that differed significantly in important ways, the data from those sites should not be aggregated together and should instead be considered separately. To control for site variance, Wells et al. (2011, 2015a) could have used a block-randomized design, as is common among multi-site trials in medicine and other fields, including Criminology (see e.g., Weisburd and Taxman 2000). Because they did not, the data have to be considered separately for the Austin and non-Austin study sites.

The Phase 1 analysis reported by Wells et al. (2015a) examined response outcomes (suspect IDs, filler IDs, and no IDs) aggregated across study sites. The data revealed a non-significant trend towards higher filler IDs for simultaneous lineups, a result that was interpreted by Wells et al. (2015a) to indicate a sequential advantage. However, when the

⁸ It is not entirely clear to us how they computed those expected values, but it is not our purpose here to dispute their validity.

data are separated by study site (Table 3A, B), as they must be given clear evidence of site variance, it becomes apparent that this non-significant trend is almost entirely a consequence of the highly conservative responding for sequential lineups in the non-Austin study sites. More conservative responding is not an indication that sequential lineups are diagnostically superior. Misinterpreting more conservative responding for diagnostic superiority is the same mistake that was made in the earlier analysis of laboratory data, where the higher diagnosticity ratio associated with sequential lineups was once thought to reflect diagnostic superiority. That result is now widely understood to instead reflect the more conservative responding often induced by the sequential procedure (National Research Council 2014).

In view of the evident site variance in the AJS field study, the Phase II analysis focusing on the “odds of guilt” associated with suspects identified from simultaneous and sequential lineups should also be examined separately by site. Our analysis of the evidentiary strength ratings data is necessarily limited to the Austin site because that is where the study was conducted. However, the case outcome data can be (and, we now realize, should be) examined separately for the Austin and non-Austin data. When that is done, the results show the trends that would be expected. The case outcome data limited to Austin (Table 4C) exhibit the expected trend favoring simultaneous lineups, ($p = .077$, Cohen’s $h = 0.511$, a medium effect size). Although not significant, this trend is expected because the expert ratings data from Austin significantly favored simultaneous lineups and because lab-based ROC data consistently favor simultaneous lineups as well. By contrast, the case outcome data limited to the non-Austin study sites (Table 4D) exhibit the expected trend that seems to favor sequential lineups. This trend is also expected because the data from the non-Austin study sites clearly show that the sequential lineups administered there (unlike in Austin) induced more conservative responding than simultaneous lineups. Under those conditions, an “odds of guilt” measure should be higher (and was higher, albeit slightly and non-significantly) for the more conservative procedure.

In response to the statistically significant expert ratings advantage for simultaneous lineups reported by Amendola and Wixted (2015a), Wells et al. (2015b) and Steblay et al. (2015) argued that those results must have been biased against the sequential procedure. The basis of their argument was that the case outcome data for the Austin sample differed noticeably from the case outcome data aggregated across all four study sites. In their view, that previously unnoticed difference exposed a hidden anti-sequential bias in the Austin sample. However, Wells et al. (2015b) and Steblay et al. (2015) were unaware of the significant variance in response bias across study sites, so they did not consider site variance as an explanation for the pattern of data that concerned them. As it turns out, site variance provides a natural explanation for what otherwise might appear to be a mysterious bias against the sequential procedure in the data reported by Amendola and Wixted (2015a).

Critically, to determine which lineup procedure is superior using any “odds of guilt” measure, the focus has to be placed on the sample that exhibited no difference in response bias across simultaneous and sequential lineups (the Austin sample), not on the sample that exhibited a clear difference in response bias across simultaneous and sequential lineups (the non-Austin sample) or on the larger sample aggregated across biased and unbiased samples. When the focus is placed on the Austin sample, the expert ratings data reveal significant advantage for the simultaneous procedure (Amendola and Wixted 2015a), and the case-outcome data reveal a marginally significant trend also favoring the simultaneous procedure (Table 4C).

Why did response bias differ between simultaneous and sequential lineups for the non-Austin study sites but not for the Austin site? We do not know. The non-Austin

eyewitnesses tested using the simultaneous procedure were more liberal than their Austin counterparts, and the non-Austin eyewitnesses tested using the sequential procedure were more conservative than their Austin counterparts (Fig. 4). It is difficult to identify a single factor that could result in opposite effects on response bias for the two lineup formats. Whether the observed site effects are attributable to undetected differences in procedure (e.g., lineup construction) or to regional differences in witness propensities to choose or to some combination of factors is unknown but is certainly worthy of further investigation.

Wells and his colleagues have been strong advocates of the sequential procedure for decades, and their strong advocacy may help to explain why, according to a national survey conducted by the Police Executive Research Forum (2013), approximately 30 % of US law enforcement agencies that use photo lineups now rely on the sequential procedure. As noted by Lindsay (1999) in an article entitled “Applying applied research: Selling the sequential line-up,” advocates of the sequential lineup procedure have relied on a variety of deliberate strategies to convince reluctant police jurisdictions to adopt the sequential procedure. Some of the recommended promotional techniques include “...initial contact with potential users, providing an incentive to adopt recommended changes, identifying and reaching decision makers, establishing trust, recognizing the importance of (even anecdotal) field data, and maintaining contact after apparent adoption of new procedures” (p. 219). In addition, the paper describes how G. Wells organized conferences and training sessions for police officers to successfully persuade them to adopt the sequential procedure. Given that the National Academy of Sciences committee on eyewitness identification just determined that the scientific evidence does not support the sequential procedure over the simultaneous procedure (National Research Council 2014), these lobbying efforts, though clearly well-intentioned, were apparently premature.

In light of their decades-long long commitment to the sequential procedure, the fact that Wells et al. (2015b) and Steblay et al. (2015) believe they have discovered evidence of an anti-sequential bias in our prior analyses of the AJS field study is perhaps understandable. However, it is essential to appreciate that, just as there is no evidence of a sequential superiority effect in lab studies that were based on the diagnosticity ratio, there is no evidence of a sequential superiority effect in the AJS field study. In both lab studies and the AJS field study, the mistake that has been made was to misinterpret the conservative responding often induced by the sequential procedure as evidence of diagnostic superiority. Conservative responding is not evidence of diagnostic superiority because (1) as noted by Clark (2012), more conservative responding simply reflects a different tradeoff between a gain (fewer false IDs of the innocent) and a cost (fewer correct IDs of the guilty); and (2) more conservative responding can be easily achieved using either lineup procedure. The superior lineup procedure is not the one that often results in more conservative responding but is instead the one that yields higher discriminability. All evidence, both in the lab and in the field to date, indicates that the simultaneous procedure yields higher discriminability than the sequential procedure.

References

- Amendola KL, Wixted JT (2015a) Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *J Exp Criminol* 11:263–284
- Amendola KL, Wixted JT (2015b) No possibility of a selection bias, but direct evidence of a simultaneous superiority effect: a reply to Wells et al. *J Exp Criminol* 11:291–294

- Amendola KL, Valdovinos MD, Hamilton EE, Slipka MG, Sigler M, 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
- Carlson CA, Carlson MA (2014) An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *J Appl Res Mem Cogn* 3:45–53
- Clark SE (2012) Costs and benefits of eyewitness identification reform: psychological science and public policy. *Perspect Psychol Sci* 7:238–259
- 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
- Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
- Gronlund SD, Carlson CA, Dailey SB, Goodsell CA (2009) Robustness of the sequential lineup advantage. *J Exp Psychol Appl* 15:140–152
- 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 Cogn* 1:221–228
- Gronlund SD, Wixted JT, Mickes L (2014) Evaluating eyewitness identification procedures using ROC analysis. *Curr Dir Psychol Sci* 23:3–10
- Innocence Project (2015) Understand the causes: the causes of wrongful conviction. New York, Innocence Project. <http://www.innocenceproject.org/causes-wrongful-conviction>. Accessed 16 March 2015
- Lindsay RCL (1999) Applying applied research: selling the sequential line-up. *Appl Cogn Psychol* 13:219–225
- Lindsay RCL, Wells GL (1985) Improving eyewitness identifications from lineups: simultaneous versus sequential lineup presentation. *J Appl Psychol* 70:556–564
- Macmillan NA, Creelman CD (2005) Detection theory: a user's guide, 2nd edn. Erlbaum, Mahwah, New Jersey
- 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
- National Research Council (2014) Identifying the culprit: assessing eyewitness identification. The National Academies Press, Washington, DC
- Police Executive Research Forum (2013) A national survey of eyewitness identification procedures in law enforcement agencies. <http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf>
- Rotello CM, Heit E, Dubé C (2015) When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychon Bull Rev* 22:944–954
- Stebly NK, Dysart J, Fulero S, Lindsay RCL (2001) Eyewitness accuracy rates in sequential and simultaneous lineup presentations: a meta-analytic comparison. *Law Hum Behav* 25:459–473
- Stebly NK, Dysart JE, Wells GL (2011) Seventy-two tests of the sequential lineup superiority effect: a meta-analysis and policy discussion. *Psychol Public Policy Law* 17:99–139
- Stebly NK, Dysart JE, Wells GL (2015) An unrepresentative sample is unrepresentative regardless of the reason: a rejoinder to Amendola and Wixted. *J Exp Criminol* 11:295–298
- Weisburd D, Taxman FS (2000) Developing a multicenter randomized trial in criminology: the case of HIDTA. *J Quant Crim* 16(3):315–340
- Wells GL, Steblay NK, Dysart JE (2011) A test of the simultaneous versus sequential lineup methods: an initial report of the AJS national eyewitness identification field studies. Des Moines, Iowa, American Judicature Society. <http://www.popcenter.org/library/reading/PDFs/lineupmethods.pdf>
- Wells GL, Steblay NK, Dysart JE (2015a) Double-blind photo-lineups using actual eyewitnesses: an experimental test of a sequential versus simultaneous lineup procedure. *Law Hum Behav* 39:1–14
- Wells GL, Steblay NK, Dysart JE (2015b) The flaw in Amendola and Wixted's conclusion on simultaneous versus sequential lineups. *J Exp Criminol* 11:285–289
- Wixted JT, Mickes L (2012) The field of eyewitness memory should abandon "probative value" and embrace receiver operating characteristic analysis. *Perspect Psychol Sci* 7:275–278
- Wixted JT, Mickes L (2014) A signal-detection-based diagnostic feature-detection model of eyewitness identification. *Psychol Rev* 121:262–276