

# Chapter 11

## Statistical Issues in Assessing the Reliability of Eyewitness Identification



Karen Kafadar

### 11.1 Introduction

Eyewitness identification (EWI) plays a critical role in criminal cases, from the investigation of the event to the courtroom trial. Almost no other evidence can be as powerful to a jury as a victim who points to a defendant and says, “Yes, he’s the one who raped me.” Yet, as we all know from our own personal experiences, memory can be fallible. The Innocence Project found that over  $\sim 70\%$  of 375+ DNA exoneration cases involved mistaken eyewitness identifications.<sup>1</sup> Wells et al. [43], Clark et al. [9] as the source of filler error rates (eyewitness mistakenly identifies a known-innocent filler as the perpetrator) is as high as 21.2% when the true perpetrator was present and 34.5% when the true perpetrator was absent. For a procedure commonly used in law enforcement proceedings, these error rates are frighteningly high. What factors lead to such high error rates? Can the levels of any of those factors be controlled by law enforcement to ensure lower error rates?

Errors in EWI can arise from: (a) identifying an innocent suspect from a lineup (*false positive*), or (b) failure to identify the correct perpetrator from a lineup when the perpetrator is present (*false negative*). Either error has severe consequences: the innocent suspect goes to jail and the true perpetrator is free to commit further crimes.

A tragic, yet not rare, example of memory infidelity arose in the case of Jennifer Thompson who, on a July evening in 1984 as a college student in North Carolina, was sexually assaulted. From Thompson’s sketch, the police collected a series of potential suspects and asked her to identify her attacker from a photo lineup. She

---

<sup>1</sup>[innocenceproject.org](http://innocenceproject.org), accessed 1 September 2021.

---

K. Kafadar (✉)

Department of Statistics, University of Virginia, Charlottesville, VA, USA  
e-mail: [kk3ab@virginia.edu](mailto:kk3ab@virginia.edu)

tentatively selected Ronald Cotton from the lineup. The police detective asked, “You’re sure?” She replied, “Positive. Did I do OK?” The detective reassured her: “You did great.” Later, Thompson was asked again to identify her attacker from a second, physical lineup of potential suspects, *in which only Ronald Cotton was included from the first lineup*. Not surprisingly, only Cotton looked familiar to her (from the previous lineup), and she confidently picked the same man. Ronald Cotton spent over 10 years in prison before he was exonerated by DNA evidence in 1995.<sup>2</sup>

The processes of encoding the event, consolidating the information in the brain, and reconstructing it at a later time, are all imperfect, leading to memory degradation and misconstrued events. When those imperfect memories lead to misidentifications, the consequences are serious, both in convicting innocent suspects and in freeing the true culprit to commit further crimes.

Despite much research on memory, relatively few factors have been considered in eyewitness identification (EWI) research, especially in view of its critical role in criminal cases. It can be the most heavily weighted evidence in a case where the available forensic evidence is either lacking (e.g., insufficient biological material for a DNA analysis) or has questionable probative value (e.g., bite marks). In view of law enforcement’s reliance on eyewitnesses, the National Academy of Sciences convened a Committee in 2013 to conduct an in-depth study of its procedures and practices, to better assess EWI error rates, and especially to identify the factors that contribute to its reliability or lead to misidentifications. The Committee issued its report in October 2014 ([31] hereafter, “NAS report”) and described aspects of vision and memory, environmental factors, and law enforcement procedures, all of which influence the accuracy of eyewitness identification. The report also discussed the use of eyewitnesses in judicial proceedings and evaluated the laboratory studies that had been conducted to date, as well as the statistical methods used to analyze the data from them and to draw conclusions from them. This chapter discusses these aspects, with particular emphasis on the design of experiments and the statistical methods that have been, and can be, used to analyze data from EWI experiments.

I chose this topic for this special volume dedicated to Steve Fienberg because of his decades-long work on statistical methods that he developed for data to ensure the equitable administration of justice, and because of his important role on the National Academy of Sciences’ Report Review Committee that led to a far better report than would have been possible without him. I will end this chapter by talking about that role, and all that I learned from him during the course of writing the NAS report.

## 11.2 Memory and Vision

While one may believe that one’s memory can be faithfully recalled like a photograph, in fact, memory is fallible and can be influenced by many factors. Memories of events involve three processes: encoding (placing aspects of the

---

<sup>2</sup>For the heartbreaking details of this case, see [www.thestory.org/stories/2013-06/jennifer-thompson](http://www.thestory.org/stories/2013-06/jennifer-thompson).

event into memory), storage (maintaining the aspects in memory), and retrieval (subsequent recall of aspects from memory). Each stage of memory is subject to degradation:

1. *Encoding*: Events may not be fully captured or encoded at the time of their occurrence due to limited extent of short-term memory, distortion and interference from effects of previous and/or subsequent events (including emotional, behavioral, sensory, observational effects), visual acuity, passage of time, and environmental conditions (e.g., poor lighting), among other factors.
2. *Storage*: The long-term retention of events can be reduced by the mere passage of time, effects of intervening experiences that replace existing memories, advancing age, and degree of associated emotion when the events occurred: “The emotional content of stored memories is a factor that appears to promote long-term retention; memories of highly arousing emotional stimuli, such as those associated with a witnessed crime, tend to be more enduring than memories of non-arousing stimuli” [31, p43].
3. *Retrieval*: Errors arise at this stage from many sources, including similarities to aspects of previously encoded and stored memories (leading to confusion between the events), loss of information about source of memory, and time. Retrieving memory of events from many decades ago may be harder for some people than those that occurred only last year; the opposite may be true for other people. Some aspects of the event may be improperly solidified into memory by the comments of others who unconsciously reinforce one’s erroneous statements about the event.

In all phases, environmental conditions and behavioral factors (e.g., stress), as well as the passage of time, can lead to further degradation in the accuracy of a witnessed event.

With all these potential effects leading to memory degradation, why is eyewitness identification even used? For some crimes, it may be the only evidence available. For some crimes, latent fingerprints cannot be captured with sufficient quality to be identifiable nor can DNA, even in trace amounts, be obtained; both can be expected to have more reliability (consistency) and greater accuracy (fewer false positives and false negatives) than EWI. Moreover, eyewitness testimony has been shown to be extremely powerful in the courtroom: jurors are highly affected by the victim who points to a defendant and says, “He is the one who attacked me. I am one hundred percent confident.” But can eyewitness identification ever be one hundred percent accurate and routinely trustworthy?

In 2014, the National Academies released a report, *Identifying the Culprit: Assessing Eyewitness Identification* [31]. The report explained features of vision and memory, and summarized much of the research that had been conducted on factors affecting the accuracy and reliability of EWI. The report also discussed the study designs that had been used for EWI experiments and the statistical methods that have been used to analyze the data from these experiments. This chapter describes some of the findings in that report, as well as subsequent studies that have been conducted. This field of inquiry demonstrates the enormous value of

cross-disciplinary research, which Steve Fienberg championed throughout his life and impressed upon his students and collaborators all over the world.

### 11.3 Challenges in Assessing EWI Reliability

The study that led to the report [31] included psychologists, legal professionals, and law enforcement personnel, as well as one lone statistician. Per the charge from the Laura and John Arnold Foundation (now Arnold Ventures), the report discussed the following issues [31, p.12].

1. Assess the existing body of scientific research related to EWI;
2. Identify research gaps in the current literature, and suggest appropriate research questions to pursue that will further our understanding of EWI and that might offer additional insight into law enforcement and courtroom practice;
3. Identify relevant research from fields outside of existing EWI research;
4. Offer recommendations to law enforcement for best practices and procedures for conducting and reporting eyewitness identifications;
5. Offer recommendations for developing jury instructions;
6. Offer advice regarding the scope of a Phase II consideration of neuroscience research and other areas of research that might affect EWI accuracy and reliability.

Briefly, the report noted these issues:

1. Much of the published literature in this field comes from researchers in university departments of psychology.
2. The existing studies have been conducted using college students or online platforms (e.g., Qualtrics<sup>®</sup>, or Amazon Mechanical Turk<sup>®</sup>); neither is likely to recreate the stress of a real-life incident. For example, the college student who must participate in an EWI experiment as a psychology course requirement, or the online participant who views a video of an attempted robbery (for a modest participation fee), is not likely to experience the same degree of anxiety and stress as would occur to a victim or witness of a real-life crime. These online platforms for conducting experiments are popular, however, because many participants can be recruited in a short period of time.
3. Most experiments in this field examine only one factor at a time, thereby preventing the exploration of interaction among factors that influence EWI reliability. Even slightly more sophisticated designs, such as factorial and fractional factorial experiments, were rarely conducted. (Admittedly, complex designs on human subjects can be more difficult to execute than on components in production environments.)
4. More statistical methods could be, but have not been, applied to analyze the data from these experiments, and thus useful data are often ignored, because researchers could not agree on how to incorporate the extra data.

This chapter describes these findings in more detail.

### 11.4 Factors that Affect EWI Accuracy

Statisticians are accustomed to approaching problems by first identifying the task at hand and then listing possible sources of variation that can affect the outcome. In this case, the task can be described as one of *binary classification*:

		Eyewitness Classification	
		“Guilty”	“Innocent”
True status of suspect	Guilty	True+	False –
	Innocent	False +	True –

In actual practice, a law enforcement officer invites the eyewitness to identify the perpetrator from a lineup, which can be conducted in several ways. Thus, the accuracy of each “binary classifier” (eyewitness) depends on many factors, including both those that can be controlled by the law enforcement officer (e.g., procedures for conducting the lineup) and those that are beyond the control of law enforcement (e.g., distance between eyewitness and the culprit at the time of the crime). In the statistics design literature, the former factors are often called “signal variables” (levels can be adjusted for maximizing the outcome) and the latter factors are “noise variables.” In the EWI literature, the terms are “system variables” (signal) and “estimator variables” (noise). Some of these factors include:

*Signal (system) variables:* Under control of law enforcement

- Protocol for lineup (e.g., live or photo; present simultaneously or sequentially)
- Number of people or photos in lineup (suspect plus “fillers”)
- Degree of similarities between suspect and fillers (“fair lineup” if very similar; “biased lineup” if suspect stands out)
- Nature of instructions to eyewitness (e.g., delivered orally or in writing; short or long)
- Presence or absence of feedback to eyewitness
- Request for confidence in identification (e.g., “How sure are you?” or “On a scale of 0–5, how confident are you?”)
- Administering official is/is not “blind” (not involved) to the circumstances of the event, including the eyewitness, crime type, location, and potential suspects.

*Noise (estimator) variables:* Beyond control of law enforcement

- Eyewitness’s level of stress or trauma at time of incident
- Conditions affecting visibility (e.g., light or dusk or dark)
- Distance between eyewitness and perpetrator
- Presence/absence of threat (e.g., weapon)
- Presence/absence of distinctive feature (e.g., scar)
- Presence/absence of other distractions (e.g., people, physical structures)
- Common/Different race or ethnicity

- Time between incident & report (*retention interval*)
- Age of eyewitness.

An important goal for law enforcement officers is to conduct EWI procedures using levels of the signal variables that maximize accuracy while minimizing variability, irrespective of the operating levels of noise variables in any given situation. Thus, it is important to design experiments that vary more than one factor at a time, especially because variables may not operate independently (Box, Hunter, Hunter 2005). However, most studies vary only one factor at a time; e.g., lineup format (sequential versus simultaneous: [23, 30]) delay between incident and the eyewitness's identification for the police officers (Deffenbacher et al. [10], Kensinger et al. [20]; Dodson et al. [12]), presence or absence of a weapon [14], same versus different race [28], process instructions to eyewitness (fair versus biased: [8]), and age [35]. Very few studies have varied multiple factors: distinctive feature and lineup format [4, 5]; distinctive feature and weapon focus [6]; lineup format, weapon focus, and distinctive feature [6]; and lineup format, weapon focus, distinctive feature, cross/same race, and target present/absent [12]. The NAS report [31] encouraged more factorial experiments.

Factorial experiments are common in several scientific fields, but they are relatively (and surprisingly) sparse in the eyewitness literature. One reason for their sparsity may lie in the need for advanced statistical methods to analyze the data from them, methods that may be unfamiliar to researchers in the field. The next section describes the most commonly used methods, followed by methods that can take account of multiple factors (which also require adjustment for multiple testing).

## 11.5 Statistical Methods: “Sequential” versus “Simultaneous”

As noted above, most studies evaluate the effect of a single factor on EWI accuracy, and often by comparing only two levels of that factor. Many of the studies were limited to considering the accuracy of identifications arising from only two levels of a single “system variable”: photo lineup format is either “sequential” versus “simultaneous;” i.e., photographs are presented to the eyewitness either one at a time or all together in a “photo array.”<sup>3</sup> In a simultaneous lineup, the witness views all individuals at the same time, and either identifies one as the perpetrator or reports that the person seen at the crime scene was not in the lineup. In a sequential procedure, the witness views individuals one at a time and reports whether or not each one is the person from the crime scene. The sequential

---

<sup>3</sup>See <http://theconversation.com/police-photo-lineups-how-background-colours-can-skew-eye-witness-identification-116329> for an example of a simultaneous lineup. The article notes the challenges in constructing a fair lineup, which include the potential bias from different background colors for all photographs and the instructions given to the eyewitness.

procedure had been in practice for decades as a possible way to reduce false identification rates, particularly after the supporting research by Lindsay and Wells [23], who “concluded that sequential presentation of lineups can reduce false identifications of innocent suspects by reducing eyewitnesses’ reliance on relative-judgment processes.” In the intervening years, an extensive literature has compared witness identification performance using simultaneous and sequential procedures. These comparisons typically have used photo-arrays.

Until recently, most studies compared the performance of simultaneous and sequential procedures in terms of a ratio of correct identifications (i.e., ‘hit rate’) to false identifications (i.e., ‘false alarm rate’), a measure known in the eyewitness identification literature as the “diagnosticity ratio” (*DR*). (The “diagnosticity ratio” is also known in other disciplines by other names; e.g., “positive likelihood ratio” or “ $LR+$  = Likelihood Ratio of a Positive Call”; see [22] Section 4.1.) The ratio corresponds roughly to confirming a positive ID; i.e., the diagnosticity ratio is related to the probability that an eyewitness who makes a positive identification correctly selected the true perpetrator. Most laboratory-based studies and systematic reviews showed that, with standard lineup instructions informing the witness that the perpetrator may or may not be present, the sequential procedure produced a higher diagnosticity ratio. That is, when a “witness” (study participant) viewed a video of a crime being committed, and later was asked to identify the culprit from a lineup, the ratio of correct to false identifications was often higher with the sequential than with the simultaneous procedure. Later research (e.g., [30]) proposed that the metric used to assess EWV accuracy was incomplete; using a different metric, simultaneous lineups led to higher accuracy. Which was correct?

A higher diagnosticity ratio (hit rate/false alarm rate) could result from a higher hit rate, a lower false alarm rate, or some combination of the two (or from some factors that affect either rate). Some early papers suggested that sequential procedures lead to fewer false alarms without changing the hit rate, which results in a higher diagnosticity ratio. More recent laboratory-based studies and systematic reviews typically show that sequential procedures are associated with a somewhat reduced hit rate accompanied by a larger reduction in the false alarm rate, thereby yielding a higher diagnosticity ratio than for the simultaneous procedure. (Results from these studies using the diagnosticity ratio have not been uniform: a recent field-based study comparing sequential to simultaneous procedures in a limited number of jurisdictions revealed a modest diagnosticity ratio advantage for the simultaneous procedure; cf. [43].) In light of this laboratory-based evidence showing a higher diagnosticity ratio for sequential lineups, many police agencies and policy makers adopted sequential lineups as their preferred EWV procedures.

However, comparisons based on a single diagnosticity ratio collapsed over all participants in that treatment arm (e.g., *DR* for sequential and *DR* for simultaneous) typically do not definitively reveal whether one procedure is “better” than the other. The EWV literature noted that a single diagnosticity ratio is affected by many factors, most importantly (a) how well a witness can *discriminate* the suspect from innocent members of the lineup and (b) by the witness’s tendency to pick or not to pick someone from the lineup; this tendency is often termed *response bias*. In general,

given the same hit rate, more conservative responding will tend to lead to fewer false alarms than more liberal responding and hence produce a higher diagnosticity ratio. Thus, a single diagnosticity ratio can be an incomplete performance measure, because it fails to account for an important variable, namely *response bias*. As noted earlier, a higher diagnosticity ratio (hit rate/false alarm rate) could result from a higher hit rate, or from a lower false alarm rate (caused by “conservative” responding), or both [29, 30, 32, 44]. The important point is that another variable (here, *response bias*) affects the measure being used to compare performance, and failure to take this variable into account can lead to naive comparisons. The receiver operating characteristic (ROC) curve was proposed as a way to account for “response bias” by using “expressed confidence level” (ECL) as a proxy for it. The ROC curve here is a plot of the *DR*’s numerator, *hit rate* (*HR*, y-axis), versus the *DR*’s denominator, *false alarm rate* (*FAR*, x-axis), calculated from the responses of participants who expressed different levels of confidence in the identification. (In general, the ROC curve is a plot of *sensitivity* versus  $(1 - \textit{specificity})$ , for different levels of a third variable; see; e.g., [2, 21].

Recognized, but often to a much lesser extent, is the fact that many other variables besides a proxy for “response bias” could easily affect sensitivity, specificity, or the diagnosticity ratio (see Sect. 11.6). Moreover, ECLs (expressed confidence levels) may have severe shortcomings as a “proxy” for “response bias” that render it insufficient for comparing two procedures, only some of which are listed below.

1. ECL is likely to be only an imperfect measure of *response bias*. A good measure of *response bias* is essential for ensuring ROC is appropriate for comparing performance of procedures. Little research has explored whether other measures would be more sensitive, and less variable, in capturing the tendency toward “conservative” versus “liberal” responding, nor how variable ECL might be for an individual presented with exactly the same circumstances and the same instructions at a later time.
2. *Discriminability* involves not only making a correct identification but also making a correct exclusion. As a plot of *hit rate* versus *false alarm rate*, the ROC curve captures the first aspect: the slope of this curve is the diagnosticity ratio at different ECLs. The second aspect involves the ratio  $\textit{specificity}/(1 - \textit{sensitivity})$ , or, using the terminology in the eyewitness literature,  $(1 - \textit{FAR})/(1 - \textit{HR})$ ; see Sect. 11.6.
3. The ECL responses in a laboratory experiment are likely to be much different in real-life, highly stressful conditions, which are very difficult (if not impossible) to replicate in an academic setting.
4. An eyewitness’s ECL of “50%” might mean something different under one procedure (e.g., “sequential”) versus another (e.g., “simultaneous”); that is, the third variable on which the ROC is based could well depend on the procedures that the ROC curves are designed to compare.
5. More than just ECL may affect the diagnosticity ratio; an ROC curve that accounts for other variables may be more useful in evaluating two procedures than an ROC curve based on only ECL (Sect. 11.6).



6. In real life, law enforcement officials recognize the impracticality of asking an eyewitness to quantify his/her stated ECL as “10%”, “20%”, . . . , “100%” and rather will invite the eyewitness to express confidence in his/her own words. The translation of a response such as “pretty confident” might indicate a “5” on a 5-point scale for one law enforcement officer but a “4” for another officer, adding to the uncertainty in ECL.
7. The use of ECL to distinguish diagnosticity ratios via an ROC curve is sensible if ECL is related to accuracy. One would hope a more confident witness is more accurate, but recent research acknowledges that the strength of this association between ECL and accuracy likely depends on many factors, such as some of the system and estimator variables already presented in the previous section (and discussed further below); see also [36].
8. The points on an ECL-based ROC curve (false alarm rate, hit rate) for respondents that expressed at least a stated level of confidence are estimates of “true” proportions; a new sample of study participants will yield a different false alarm rate, different hit rate, and different number of participants in that ECL category. The uncertainties in these proportions are rarely displayed, but they should be, and they can indicate different conclusions; see below.
9. Finally, ROC curves are often constructed by plotting the “hit rate” versus “false alarm rate” at different levels of ECLs *across all subjects in the study*. In laboratory experiments where the majority of eyewitness identification studies have been conducted (and sometimes in field studies), the same subject might be asked to view more than one condition, responding “10% confident” in one condition but “50% confident” in another condition. The effects of this lack of independence among subjects on the bias and variability in the estimated ROC curve (which requires estimates of “hit rate” = sensitivity and “false alarm rate” = 1 – specificity) have yet to be quantified. Moreover, the plotted points themselves have uncertainty (see below).

Incidentally, the procedures governing the administration of “sequential lineups” that are used in laboratory settings may differ in substantive ways from those in real life. For example, in a lab setting, the “eyewitness” sees each photo only once and cannot “go back” to previous photos. Conversely, in a real-life setting, a real eyewitness may ask to see a previously viewed photo, and the law enforcement officer (who is administering the lineup) often will allow him/her to go back and view previously shown photographs. This discrepancy in the procedures between lab and real life is one of the reasons why the generalization of inferences from lab experiments to real-life practice may be questionable. The inability to faithfully replicate in a lab the stress of the scene on the eyewitness is another reason for the disconnect between lab experiments and real-life incidents.

#### *Potential Uncertainties in ECL-based ROC Curve*

The construction of the ECL-based ROC curve requires study participants to answer after each of their choices, “How confident are you in your decision?” Examples of scales for their responses are (1) 11-point scale “0%, 10%, 20%, . . . , 100%”, (2) 7-point scale of 1 (not confident at all) to 7 (highly confident),

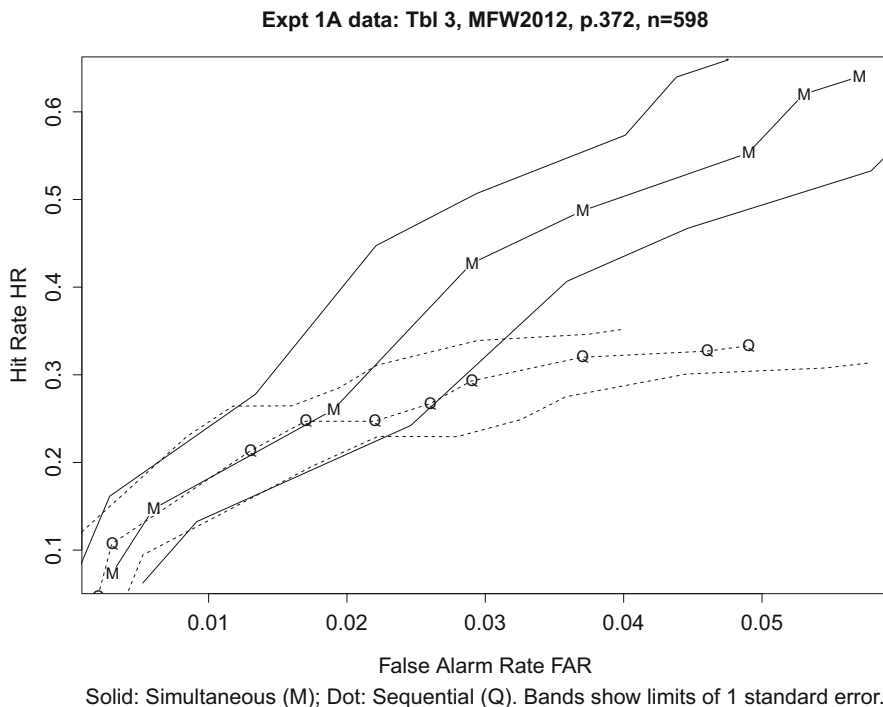
or (3) 6-point scale “0%, 20%, 40%, 60%, 80%, 100%”. For example, Brewer and Wells ([3], Table 9, p.24) conducted an experiment with 1200 recruited participants (undergraduates and from the local community) who watched a video of a thief in a restaurant and later were asked to identify the thief from a simultaneous lineup of eight color photographs. Roughly 300 participants viewed a lineup with a “foil” that had either high or low similarity to the culprit, and where the true culprit was either present or absent.<sup>4</sup> They calculated different DRs depending on the ECL category: 3.5 (“0–20%”), 3.3 (“30–40%”), 5.9 (“50–60%”), 13.6 (“70–80%”), 38.3 (“90–100%”), for the 524 participants who chose a suspect from the lineup. (Interestingly, the DRs change little across these five ECL categories for the 676 participants who declined to choose any suspect: 1.0, 0.9, 1.2, 1.8, 2.3, respectively.) Mickes et al. [30, Table 1, p.367] then calculated the DRs using slightly different categories and found that they varied, mostly in the last (90–100%) category: 9.0 (“< 0%”), 9.3 (“< 30”), 10.7 (“< 50%”), 16.0 (“< 70%”), 43.2 (“< 90%”). Similar to the use of receiver operating characteristics (ROCs) used in diagnostic medicine to compare different diagnostic test modalities, these authors argued that, by plotting the hit rate (sensitivity) versus the false alarm rate (1 – specificity) collapsed over participants *who express a given ECL or higher*, the area under the resulting ROC curve (AUC) provides a more appropriate metric for comparison. In doing so, a comparison of AUCs in experiments comparing sequential (seq) and simultaneous (sim) lineups,  $AUC(sim) > AUC(seq)$ . Notice that the slope of the ROC curve at the plotted point is the DR for that ECL category (“at least x% confident”); hence, a straight line would indicate no effect of ECL on DR.

Figure 11.1 shows data published from a similar (thief) experiment in Mickes et al. ([30] “Experiment 1a” in Table 3) that plots the hit rate (*HR*) versus the false alarm rate (*FAR*) for sequential (Q) versus simultaneous (M) lineups. Their plot (Fig. 11.6a, p.371) has been enhanced here in Fig. 11.1 by the addition of two curves that represent very optimistic standard errors on the data points (binomial variation in the reported *HR* and *FAR*). Indeed, simultaneous (M) appears to have a higher AUC than sequential (Q), especially at higher levels of *HR* and *FAR*. Data from another experiment analyzed in Liu [26] also indicated the possible superiority of simultaneous over sequential. But data from other experiments are far less clear: Fig. 11.2 shows the same type of curve, for data from “Experiment 2” in Mickes et al. [30, Table 3]. Suddenly, now, the advantages of simultaneous (M) over sequential (Q) are not so clear.

#### *Does Higher Confidence Lead to Greater Accuracy?*

The underlying premise of the ROC approach over the single collapsed *DR* is based on the theory that increased accuracy is associated with increased confidence in the eyewitness’s identification. (Recall from paragraph 2 of this section that *DR*

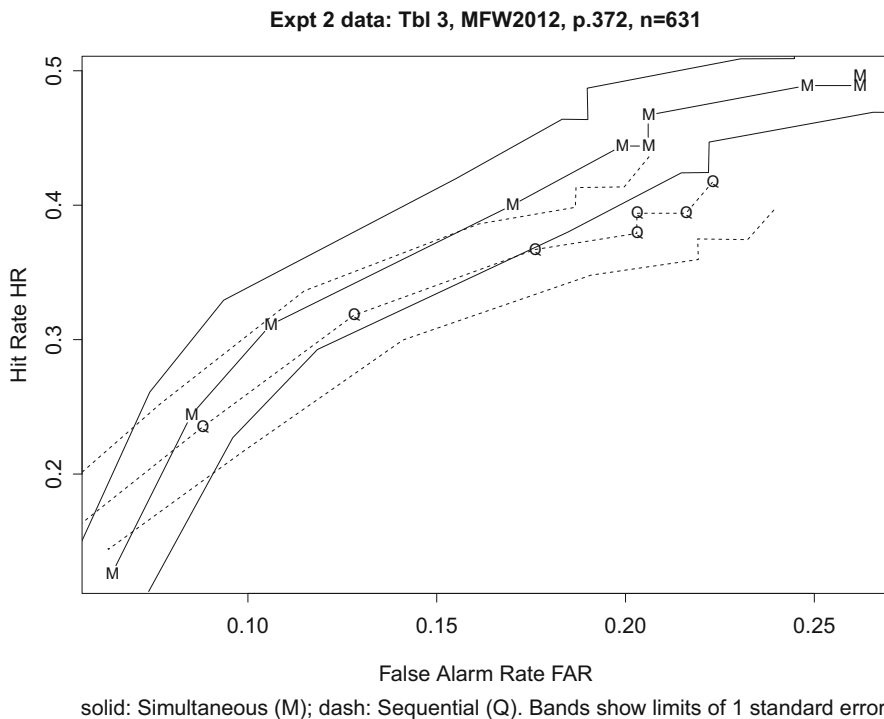
<sup>4</sup>The actual counts in Table 2 in Brewer and Wells [3] are: 299 (high similarity, target present); 300 (low similarity, target present); 301 (high similarity, target absent); 300 (low similarity, target absent). In each of these four conditions, ~150 participants viewed a “biased” lineup and the other half viewed an “unbiased” (sometimes called “fair”) lineup.



**Fig. 11.1** “ROC curve” (hit rate = sensitivity versus false alarm rate =  $1 - \text{specificity}$ ), using data from Experiment 1A in [30, p.372]. Based on 598 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated HR and FAR rates. Points on ROC curve refer to “At least 10% confident,” . . . , “At least 90% confident,” “At least 100% confident”. Figure C-3 from [31], p.148: Reproduced with permission from the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

is the same as the “positive likelihood ratio”  $LR+$ .) Data from several experiments suggest that such an association may exist. For example, Figs. 11.3 and 11.4 are plots of DR versus ECL using the same data from [30] as in Figs. 11.1 (“Experiment 1a”) and 11.2 (“Experiment 2”) above, respectively; see also [11]. However, studies are not consistent in this relationship; e.g., see Juslin et al. [19]. Further, given the multiple sources of variation that can affect ECL, the association is likely to vary greatly depending on the eyewitness, conditions of the event, and other factors yet to be identified.

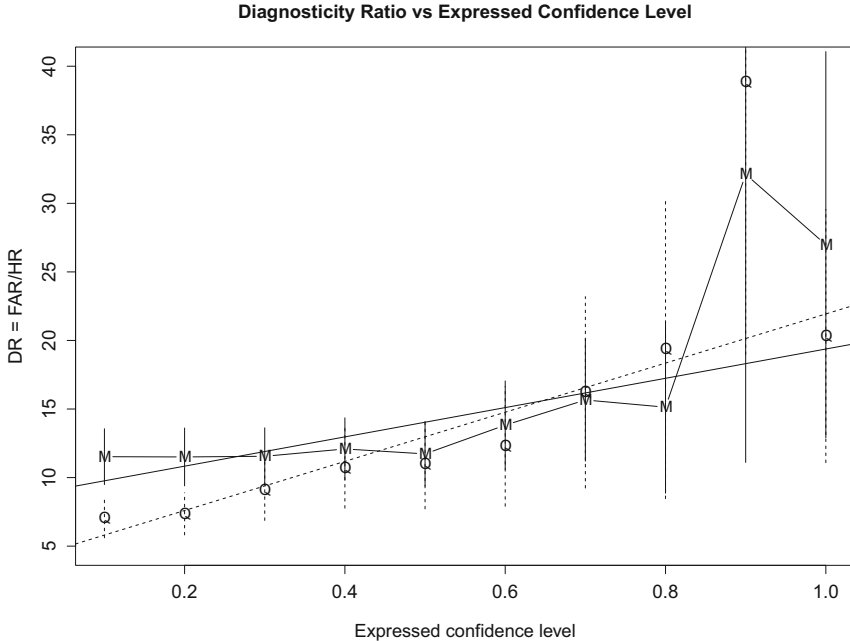
Thus, even the migration from a single DR to an ECL-based ROC may not adequately capture all the factors that should be taken into account when comparing two procedures (such as “simultaneous” versus “sequential” lineups). Incidentally, few of the articles in the EWI literature take account of multiplicity in hypothesis testing: adjustment for multiplicity, such as the false discovery rate [1], appears to be very rare.



**Fig. 11.2** “ROC curve” (hit rate = sensitivity versus false alarm rate =  $1 - \text{specificity}$ ), using data from Experiment 2 in [30, p.372]. Based on 556 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated HR and FAR rates. Points on ROC curve refer to “At least 10% confident,” . . . , “At least 90% confident,” “At least 100% confident”. Figure C-4 from [31], p.140: Reproduced with permission from the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

## 11.6 Improved Statistical Approaches: PPV and NPV

The primary goal of an eyewitness procedure is to maximize *both* the *Positive Predictive Value (PPV)* as well as the *Negative Predictive Value (NPV)*. Neither can be observed in real life. We *can* estimate two of the components in the definitions of PPV and NPV, namely *sensitivity* and *specificity*. *Sensitivity* is defined as the probability that an eyewitness, who is shown the *true* culprit, correctly identifies him/her as the culprit. *Specificity* is defined as the probability that an eyewitness, who is shown the *true* innocent suspect, correctly does not identify him/her as the culprit. Both probabilities can be estimated only in a designed study, where the study facilitators *know* the true status of the culprit and innocent suspects. In real life, one does not know the true status. At best, we can only estimate the probability that a positive ID correctly identified the true culprit (“positive predictive



Data from MFW2012, p.372, Expt 1A: M=Simultaneous (solid), Q=Sequential (dash); limits of 1 standard error

**Fig. 11.3** Plot of DR = diagnosticity ratio = hit rate (sensitivity) divided by false alarm rate (1 – specificity), using data from Experiment 1a in [30, p.372]. Based on 598 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated HR and FAR rates

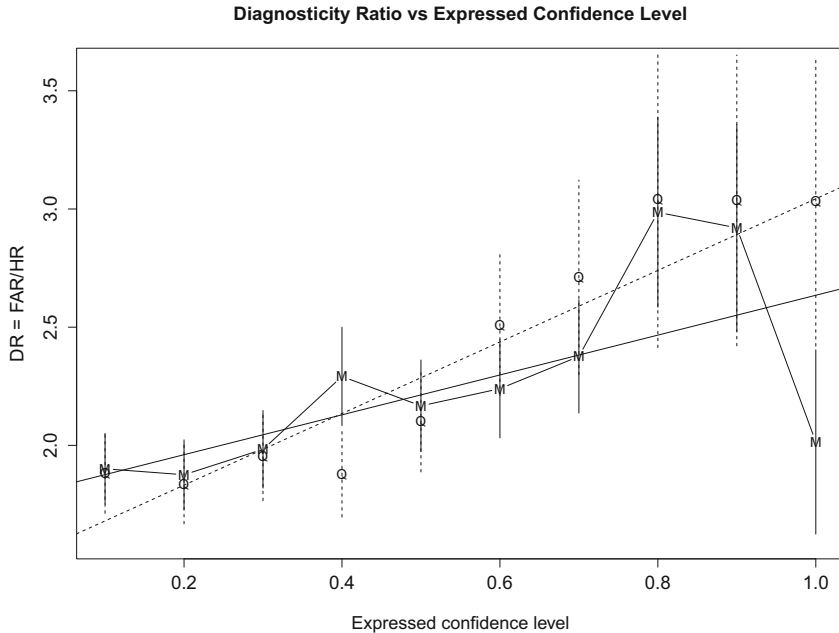
value”, or PPV) and the probability that no ID correctly recognized an innocent suspect (“negative predictive value”, or NPV). Both PPV and NPV depend on three quantities (sensitivity; specificity; and prevalence, or the proportion of individuals that might be the culprit; i.e., 1 out of 6, or 1 out of 100, or ...), but in different ways, as explained below.

Let  $S$  denote *sensitivity*, or “hit rate,”  $T$  denote *specificity*, or “1 – false alarm rate,” and  $p$  denote the probability that the suspect is the perpetrator (*prevalence*). Let  $OR$  denote the “odds ratio” =  $(1 - p)/p$ . Then

$$PPV = (Sp)/[Sp + (1 - T)(1 - p)] = 1/(1 + OR/d), d = S/(1 - T)$$

$$NPV = [T(1 - p)]/[T(1 - p) + (1 - S)p] = 1/[1 + ((1 - S)/T)/OR].$$

Recall that  $S/(1 - T)$  = (hit rate/false alarm rate) is the same as the diagnosticity ratio ( $DR$ ), or, more generally,  $LR+$  = “Likelihood Ratio of a Positive Call;” it measures roughly the probability of a correct ID. So method 1 is preferred over method 2 if  $PPV_1 > PPV_2$ ; i.e.,  $(OR_1/DR_1) < (OR_2/DR_2)$ ; i.e.,  $(DR_1/OR_1) > (DR_2/OR_2)$ , where  $DR_1, DR_2$  denote the diagnosticity ratios for methods 1 and

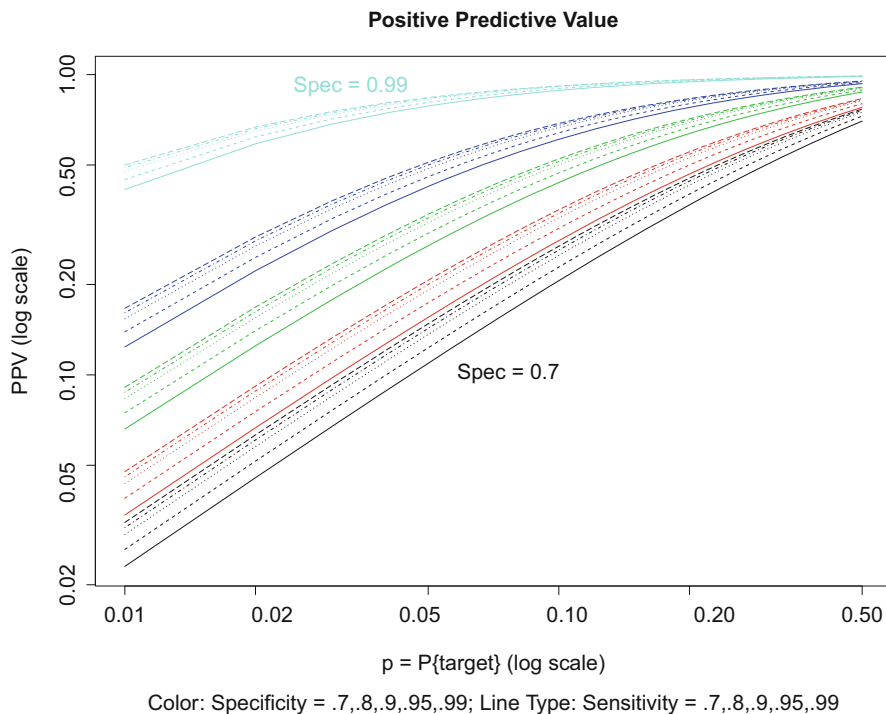


Data from MFW2012, p.372, Expt 2: M=Simultaneous (solid); Q=Sequential (dash); limits of 1 standard error

**Fig. 11.4** Plot of DR = diagnosticity ratio = hit rate (sensitivity) divided by false alarm rate (1 – specificity), using data from Experiment 2 in [30, p.372]. Based on 556 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated HR and FAR rates

2, respectively, under the same circumstances (i.e.,  $OR_1 = OR_2$ , same conditions for both methods, etc.). Thus, a comparison of procedures in terms of criteria based on the diagnosticity ratio ( $LR+$ ) is a comparison in terms of  $PPV$ , the probability that an eyewitness’s ID really identified the true culprit.

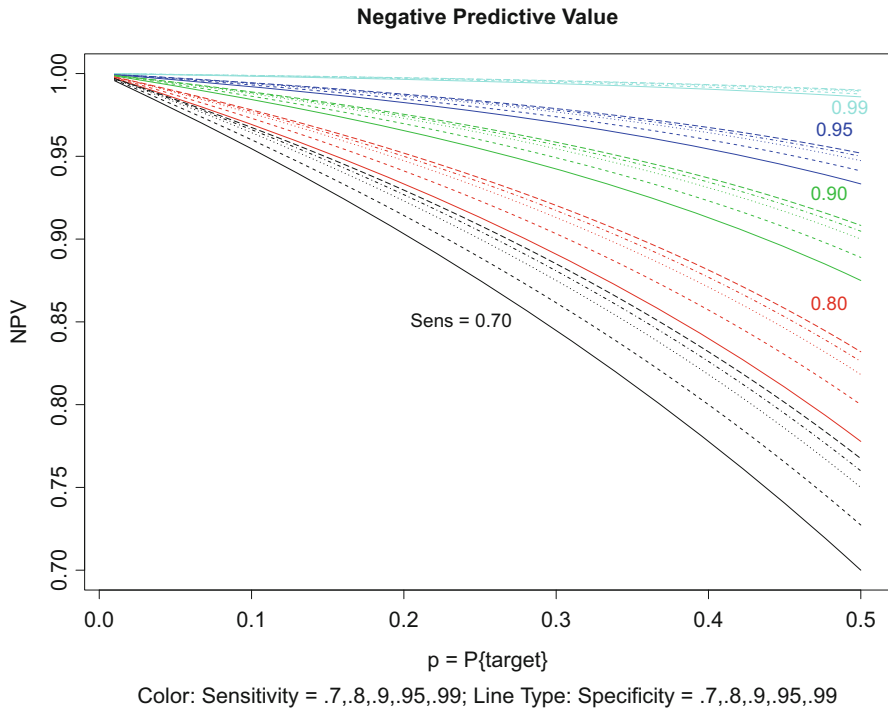
To take into consideration the second aspect, NPV (probability that a non-ID correctly recognized an innocent suspect in the lineup), a second ratio is needed. Corresponding to  $LR+ = S/(1 - T)$ , this ratio,  $(1 - S)/T$ , is known as  $LR- =$  “Likelihood Ratio of a Negative Call;” it corresponds roughly to the likelihood of ruling out a suspect ([22] §4.1). A calculation for NPV, similar to the one above for  $PPV$ , shows that method 1 is preferred over method 2 if  $NPV_1 > NPV_2$ ; i.e.,  $T_1/(1 - S_1) > T_2/(1 - S_2) = 1/(LR-)_1 > 1/(LR-)_2$  – again, assuming that the odds ratio  $(1 - p)/p$  stays constant in the comparison between the two procedures (e.g., both the simultaneous and the sequential lineups have the same number of “foils”). Both the single diagnosticity ratio and the ROC curve (which shows separate diagnosticity ratios as slopes of the curve at the points corresponding to different levels of expressed confidence) address the  $PPV$  criterion; a separate analysis is needed for the NPV criterion. (Note that all calculations have variability,



**Fig. 11.5** Plot of Positive Predictive Value (PPV) versus prevalence, for levels of specificity (black = 0.7, red = 0.8, green = 0.9, blue = 0.95, turquoise = 0.99) and sensitivity (solid = 0.7, short dash = 0.8, dot = 0.9, dot-dash = 0.95, long dash = 0.99). Both PPV and prevalence are on logarithmic scales. Note that specificity (color) affects PPV much more than sensitivity (line type), and that both high specificity and high prevalence are necessary for high PPV

due to uncertainties in estimating  $S$  and  $T$ .) The effects of sensitivity, specificity, and prevalence on PPV and NPV are shown in Figs. 11.5 and 11.6, respectively. The figures show that *specificity* affects PPV more than *sensitivity*, that *sensitivity* affects NPV more than *specificity*, and that very high values of both sensitivity and specificity, as well as very high prevalence, are needed for high PPV and NPV.

Thus, although the ECL-based ROC curve is an “improvement” over the single DR collapsed over all subjects, in that it takes into account the eyewitness’s expressed confidence in the ID (often seen as a proxy for “response bias” and shown in some studies to be correlated to some degree with accuracy), it focuses on only PPV and ignores NPV as well as other variables that could be affecting accuracy. For sufficiently large experiments, one may be able to construct ROC curves for each participant in the study, characterizing the information in each curve by a summary measure (such as logarithm of the Area Under the Curve; bigger is better) as a function of several factors. For example, [42] propose a hierarchical model for  $\log(AUC)$  to distinguish within-participant variability from between-



**Fig. 11.6** Plot of Negative Predictive Value (NPV) versus prevalence, for levels of sensitivity (black = 0.7, red = 0.8, green = 0.9, blue = 0.95, turquoise = 0.99) and specificity (solid = 0.7, short dash = 0.8, dot = 0.9, dot-dash = 0.95, long dash = 0.99). Both PPV and prevalence are on logarithmic scales. Note that sensitivity (color) affects NPV much more than specificity (line type), and that both high sensitivity and low prevalence are necessary for high NPV

participant variability; other effects can be incorporated into the model.<sup>5</sup> Shiu and Gatonis [37] offer a methodology for incorporating both  $LR+$  and  $LR-$  through the “predictor receiver operating characteristic” (PROC) curve, by plotting PPV versus  $1 - NPV$ . Other researchers have suggested approaches that would be even more useful, namely bivariate logistic regression models for logit-transformed sensitivity and specificity [34] generalized by Chu and Cole [7]. DuMouchel [13], Junaidi and Stojanovski [18], and Luby [27] also propose hierarchical models that are applicable for data from EWI experiments. See Liu et al. [25, Section 21.4] for further details about using these models in this context.

As noted at the start of Sect. 11.4, the basic task of the eyewitness is to serve as a *binary classifier*. The statistics and computer science literature contains many

<sup>5</sup>Notice that Wang and Gatsonis use AUC, versus pAUC = partial area under the curve; see discussions about AUC versus pAUC in the articles by Pepe [33, p.311] and Walter [40].



proposals for combining the information from binary classifiers; e.g., see [17]. Liu et al. [25, Section 21.4] successfully use random forests and mention other supervised methods that can be applied to the comparison of EWI procedures in the presence of multiple factors, including the “noise” (“estimator”) variables that were mentioned above (lighting, distance, race, etc.). Probably the simplest approach would be ordinary logistic regression, which provides a model for the probability of being accurate in light of other manipulated variables [15]; see also Appendix C in the [31] report (pages 150–154) for an analysis of data from an experiment conducted by Carlson and Carlson [6]. The research in the EWI field is yet another example of the benefits of cross-disciplinary research involving statisticians.

## 11.7 Conclusion and Postscript

This article describes only briefly the state of the experimental designs and statistical methods that have been used in the literature on assessing the reliability of eyewitness identification procedures. The potential for statisticians to contribute to this field is immense, but not without its challenges.

I chose this topic for this chapter because I encountered these challenges while serving as a member of the Committee that authored the [31] report. Steve Fienberg served as co-chair of the Report Review Committee that was in charge of ensuring “that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered” [31, p.xii]. Despite multiple obligations on his plate at the time, Steve was masterful in his role. At one point, the Committee requested a call with him and his co-Chair, Dr. David Korn. With no advanced warning of the Committee’s queries to them, Steve responded thoughtfully, eloquently, and precisely to each question raised by Committee members regarding various aspects of the review. His explanations of statistical concepts were crystal clear, emphasizing their complexities and relevance in an unambiguous but non-condescending manner. As I listened on the phone, I was wishing I could have recorded the call; it was impressive, even if he had been forewarned, but nothing short of spectacular in view of the fact that he had no idea what he would be asked. I always had great admiration for Steve from the time I met him, but that call impressed me immeasurably. And so it is, whenever I think of the NRC report, I remember with great fondness the influence that Steve Fienberg had on it, in ways that no one would ever know.

**Acknowledgments** An early version of this paper was presented at a workshop during the Probability and Statistics in Forensic Science program at the Isaac Newton Institute for Mathematical Sciences supported by EPSRC Grant Number EP/K032208/1. The author thanks the Institute for its hospitality and support during the program. The author also gratefully acknowledges her co-investigators on a grant from Arnold Ventures (Chad Dodson, Brandon Garrett, Joanne Yaffe). This chapter was prepared in part with support from this grant. The views expressed herein are solely those of the author and do not necessarily represent the views of Arnold Ventures or the Isaac Newton Institute.

## References

1. Benjamini, Yoav; Hochberg, Yusef (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* 57: 289–300.
2. Brown, Christopher D.; Davis, Herbert T. (2006), Receiver operating characteristics curves and related decision measures: A Tutorial, *Chemometrics and Intelligent Laboratory Systems* 80: 24–38.
3. Brewer, Neil; Wells, Gary L. (2006). The Confidence-Accuracy Relationship in Eyewitness Identification: Effects of Lineup Instructions, Foil Similarity, and Target-Absent Base Rates,” *Journal of Experimental Psychology: Applied* 12(1): 11–30.
4. 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(2): 118–128.
5. Carlson, C.A.; (2011). Influence of a perpetrator’s distinctive facial feature on eyewitness identification from simultaneous versus sequential lineups, *Applied Psychology in Criminal Justice* 7:77–92.
6. Carlson, Curt A.; Carlson, Maria A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis, *Journal of Applied Research in Memory and Cognition* 3(2): 45–53.
7. Chu, Haitao; Cole, Stephen R. (2006), Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach, *Journal of Clinical Epidemiology* 59:1331–1333.
8. Clark, Steven E. (2005), A re-examination of the effects of biased lineup instructions in eyewitness identification, *Law and Human Behavior* 29(4): 395–424.
9. Clark, Steven E.; Howell, Ryan T.; Davey, Sherrie L. (2008), Regularities in eyewitness identification. *Law and Human Behavior*, 32(3), 187–218. <https://doi.org/10.1007/s10979-006-9082-4>
10. Deffenbacher K.A.; Bornstein, B.H.; McGorty, E.K.; Penrod, S.D. (2008), Forgetting the once-seen face: estimating the strength of an eyewitness’s memory representation, *Journal of Experimental Psychology: Applied* 14(2): 139–150.
11. Dodson, Chad S.; Dobolyi, David G. (2016), Confidence and eyewitness identifications: the cross-race effect, decision time, and accuracy. *Applied Cognitive Psychology*, 30:113–125.
12. Dodson, Chad S.; Garrett, Brandon L.; Kafadar, Karen; Yaffe, Joanne (2020), Eyewitness Identification Speed: Slow identifications from highly confident eyewitnesses hurt perceptions of their testimony, *Journal of Applied Research in Memory and Cognition* 10:259–267.
13. DuMouchel, William (1994), Hierarchical Bayes linear models for meta-analysis. Technical Report 27, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, PO Box 14006, Research Triangle Park, NC 27709, September 1994.
14. Fawcett, J.M.; Russell, E.J.; Peace, K.A.; Christie, J. (2013), Of guns and geese: A meta-analytic review of the ‘weapon focus’ literature, *Psychology, Crime & Law* 19:35–66. <https://doi.org/10.1080/1068316X.2011.599325>
15. Garrett, Brandon L.; Liu, Alice J.; Kafadar, Karen; Yaffe, Joanne; Dodson, Chad S. (2020): Factoring the Role of Eyewitness Evidence in the Courtroom. *Journal of Empirical Legal Studies* 17(3): 556–579.
16. GEP Box, WG Hunter, JS Hunter (2005), *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed., Wiley, New York.
17. Hastie, Trevor; Friedman, Jerome H.; Tibshirani, Robert (2013), *The Elements of Statistical Machine Learning*, 2nd edition, Springer: New York.
18. Junaidi, Darfiana Nur; Stojanovski, Elizabeth (2012), Bayesian estimation of a meta-analysis model using Gibbs sampler. In *Proceedings of the Fifth Annual ASEARC Conference - Looking to the future - Programme and Proceedings*.

19. Juslin, Peter; Olsson, Nils; Winman, Anders (1996), Calibration and diagnosticity of confidence in eyewitness identification: comments on what can be inferred from low confidence-accuracy correlation. *Journal of Experimental Psychology*, 22(5):1304–1316.
20. Kensinger, E.A.; Schacter, D.L. (2008), Neural processes supporting young and older adults' emotional memories. *Journal of Cognitive Neuroscience* 20(7):1161–1173.
21. Lasko, Thomas A.; Bhagwat, Jui G.; Zou, Kelly H; Ohno-Machado, Lucila (2005), The use of receiver operating characteristic curves in biomedical informatics, *Journal of Biomedical Informatics* 38(5), 404–415.
22. Lee, Peter (2012), *Bayesian Statistics: An Introduction, 4th ed.*, Wiley, New York.
23. Lindsay, R.C.L., Wells, Gary L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70(3), 556–564. <https://doi.org/10.1037/0021-9010.70.3.556>
24. Lindsay, R.C.L. (1999), Applying applied research: Selling the sequential lineup, *Applied Cognitive Psychology*, 13, 219–225.
25. Liu, Alice; Kafadar, Karen; Garrett, Brandon; Yaffe, Joanne (2020), Bringing new statistical approaches to eyewitness evidence. Chapter 21 in *Handbook of Statistics in Forensic Science* (Eds. David L. Banks, Karen Kafadar, David H. Kaye, Maria Tackett), CRC Press.
26. Liu, Alice Jia (2020). Errors-In-Variables and Random Forests: Theory and Application to Eyewitness Identification Data. Ph.D. Dissertation, University of Virginia, Charlottesville, Virginia
27. Luby, Amanda S. (2017), Strengthening analyses of line-up procedures: a log-linear model framework, *Law, Probability and Risk* 16:241–257.
28. Meissner, C.A.; Brigham, J.C. (2001), Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review, *Psychology, Public Policy, and Law* 7(1): 3–35, <https://doi.org/10.1037/1076-8971.7.1.3>
29. Meissner, C.A.; Tredoux, C.G.; Parker, J.F.; MacLin, O.H. (2005), Eyewitness decisions in simultaneous and sequential lineups: A dual process signal detection theory analysis, *Memory & Cognition* 33, 783–792.
30. Mickes, Laura; Flowe, Heather D.; Wixted, John T. (2012), Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups, *Journal of Experimental Psychology: Applied*, 18(4): 361–376.
31. National Research Council (NRC 2014): *Identifying the Culprit: Assessing Eyewitness Evidence*, National Academies Press (<https://doi.org/10.17226/18891>).
32. Palmer, M.A., Brewer, N. (2012), Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior*, 36, 247–255.
33. Pepe, Margaret Sullivan (2000), Receiver Operator Characteristic Methodology, *Journal of the American Statistical Association* 95(449), 308–311.
34. Reitsma, Johannes B.; Glas, Afina S.; Rutjes, Anne W.S.; Scholten, Rob J.P.M.; Bossuyt, Patrick M.; Zwinderman, Aeilko H. (2005), Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58:982–990, 2005.
35. Rhodes, M.G.; Anastasi, J.S. (2012), The Own-Age Bias in Face Recognition: A Meta-Analytic and Theoretical Review,” *Psychological Bulletin* 138(1): 146–174.
36. Roediger, H.L.; Wixted, J.H.; DeSoto, K.A. (2012), The curious complexity between confidence and accuracy in reports from memory, in: L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and Law*, Oxford, Oxford University Press, 84–118.
37. Shiu, Shang-Ying; Gatsonis, Constantine A. (2008), The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values, *Philosophical Transactions of the Royal Society A* 366:2313–2333.
38. Steblay, Nancy K.; Dysart, Jennifer; Fulero, Solomon; Lindsay, R.C.L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison, *Law and Human Behavior* 25(5): 459–473

39. Steblay, Nancy K.; Dysart, Jennifer; Wells, Gary L. (2011). Seventy-two Tests of the Sequential Lineup Superiority Effect: A Meta-Analysis and Policy Discussion, *Psychology Public Policy and Law* 17(1): 99–139,
40. Walter, S.D. (2005), The partial area under the summary ROC curve, *Statistics in Medicine* 24(13), 2025–2040.
41. Wells, Gary L.; Rydell, S.M.; Seelan, E.P. (1993), The selection of distractors for eyewitness lineups, *Journal of Applied Psychology*, 78, 835–844.
42. Wang, Pei; Gatsonis, Constantine (2008), Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests, *Statistics in Medicine* 27:243–256 (<https://doi.org/10.1002/sim.2828>).
43. Wells, Gary L.; Steblay, Nancy K.; Dysart, Jennifer E. (2015), Double-Blind Photo-Lineups Using Actual Eyewitnesses: An Experimental Test of a Sequential versus Simultaneous Lineup Procedure, *Law and Human Behavior* 39(1), 1–14. <https://doi.org/10.1037/lhb0000096>.
44. Wells, Gary L. (2014) Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup, *Current Directions in Psychological Science*, 23:11–16.
45. Wells, Gary L.; Kovera, Margaret B.; Douglass, Amy B.; Brewer, Neil, Meissner, Christian A.; Wixted, John T. (2020), Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence, *Law and Human Behavior*, 44(1), 3–36. <https://doi.org/10.1037/lhb0000359>