Law and Human Behavior
2017, Vol. 41, No. 2, 127–145

0147-7307/17/$12.00 http://dx.doi.org/10.1037/lhb0000219

Fair Lineups Are Better Than Biased Lineups and Showups, but Not Because They Increase Underlying Discriminability

Andrew M. Smith and Gary L. Wells
Iowa State University

R. C. L. Lindsay
Queen's University

Steven D. Penrod
John Jay College of Criminal Justice, City University of New York

Receiver Operating Characteristic (ROC) analysis has recently come in vogue for assessing the underlying discriminability and the applied utility of lineup procedures. Two primary assumptions underlie recommendations that ROC analysis be used to assess the applied utility of lineup procedures: (a) ROC analysis of lineups measures underlying discriminability, and (b) the procedure that produces superior underlying discriminability produces superior applied utility. These same assumptions underlie a recently derived diagnostic-feature detection theory, a theory of discriminability, intended to explain recent patterns observed in ROC comparisons of lineups. We demonstrate, however, that these assumptions are incorrect when ROC analysis is applied to lineups. We also demonstrate that a structural phenomenon of lineups, differential filler siphoning, and not the psychological phenomenon of diagnostic-feature detection, explains why lineups are superior to showups and why fair lineups are superior to biased lineups. In the process of our proofs, we show that computational simulations have assumed, unrealistically, that all witnesses share exactly the same decision criteria. When criterial variance is included in computational models, differential filler siphoning emerges. The result proves dissociation between ROC curves and underlying discriminability: Higher ROC curves for lineups than for showups and for fair than for biased lineups despite no increase in underlying discriminability.

Keywords: eyewitness identification, lineups, Receiver Operating Characteristic (ROC) analysis, criterial variance, Signal Detection Theory

A lineup is a procedure in which a single suspect is embedded among known innocents (hereafter fillers) and presented to an eyewitness to test law enforcement personnel's hypothesis that the suspect is the culprit. The suspect might be the culprit, but it is also possible that the suspect is innocent. The rationale for using a lineup is to gain information about the likelihood that the suspect is the culprit. An identification of the suspect increases the probability that the suspect is the culprit; an identification of a filler or rejection of the lineup decreases the probability that the suspect is the culprit (Wells & Lindsay, 1980; Wells, Yang, & Smalarz, 2015). Though law enforcement personnel do not know if a suspect is guilty (the suspect could be innocent), eyewitness researchers routinely run experimental studies in which they systematically manipulate the presence and absence of a to-be-remembered individual (hereafter culprit) in lineup procedures. When the culprit is present in a lineup the eyewitness can answer correctly by identifying the culprit or incorrectly by rejecting the lineup or identifying a filler. When the culprit is absent the eyewitness can answer correctly by rejecting the lineup or incorrectly by identifying the innocent suspect or a filler.

In recent research, Signal Detection Theory (SDT, a much-used tool in the study of recognition memory; Egan, 1958; Green & Swets, 1966) has been used to analyze lineup identification procedures as if the lineup paradigm and the traditional recognition paradigm were identical. However, we argue that the lineup problem is distinct from the classic 2 (affirmative response vs. negative response) × 2 (signal plus noise vs. noise alone) problem structure to which SDT is commonly applied in recognition memory. Indeed, the lineup procedure has a 3 (suspect identification, filler identification, rejection) × 2 (culprit present vs. culprit absent) problem structure and it is important for both our theoretical understanding of how lineups work and for assessing their applied utility that each of these three categories of responses (suspect identifications, filler identifications, and rejections) is examined. This position is in direct contrast to the recently advocated approach of using Receiver Operating Characteristic (ROC) analysis to make theoretical and applied inferences about lineups, which proposes that researchers only need to examine suspect identifications (Wixted & Mickes, 2012).

The presumption behind ROC analyses for lineups is that ROC curves reflect underlying discriminability (Wixted &

This article was published Online First September 29, 2016.

Andrew M. Smith and Gary L. Wells, Department of Psychology, Iowa State University; R. C. L. Lindsay, Department of Psychology, Queen's University; Steven D. Penrod, Department of Psychology, John Jay College of Criminal Justice, City University of New York.

Correspondence concerning this article should be addressed to Andrew M. Smith, Department of Psychology, Iowa State University, W112 Lagomarcino Hall, Ames, IA 50011. E-mail: amsmith@iastate.edu

Mickes, 2015a, 2015b). Underlying discriminability refers to the capacity of a procedure to help witnesses sort between novel faces and previously seen faces. But we will show that ROC analyses, when applied to lineups, do not reflect underlying discriminability. A second claim is that the identification procedure that produces better underlying discriminability is the procedure that should be preferred for applied purposes. However, using two of the most fundamental problems in eyewitness identification, we will show a dissociation between the underlying discriminability of a procedure and the applied value of that procedure. The prominent examples that we use are lineups versus showups and fair lineups versus biased lineups. We show that fair lineups actually harm underlying discriminability relative to showups or biased lineups but nevertheless produce better applied outcomes and we explain why that is the case.

We differentiate between three distinct diagnostic properties of lineup identification procedures: underlying discriminability, positive predictive value (PPV), and negative predictive value (NPV). As already noted, underlying discriminability refers to the capacity of a procedure to facilitate eyewitnesses in sorting between novel faces and a previously seen face. However, when calculating discriminability it is important to note that novel faces include not just the face of an innocent suspect but also the faces of innocent fillers in both the culprit-present and culprit-absent lineups. PPV, in contrast, refers to the capacity of a procedure to incriminate a suspect and is calculated using only identifications of guilty and innocent suspects (not fillers). ROC analysis on lineups is a measure of PPV, but it is not the only measure of PPV. NPV refers to the capacity of a procedure to exculpate a suspect and is derived using calculations on identifications of fillers and calculations on rejection decisions. ROC analyses do not measure NPV.

Because our focus in the current work is on examining the utility of ROC analysis on lineups, we do not examine NPV in the current work. Instead we focus on what ROC analysis on lineups is assumed to measure, namely, PPV and underlying discriminability. By no means are we attempting to dismiss the importance of NPV from lineup procedures. The National Research Council (2014) recently discussed the importance of examining the NPV of lineup procedures and we could not agree more with this recommendation. NPV is an important, but often overlooked, property of identification procedures. Indeed, some forensic tests are valuable primarily for their NPV rather than for their PPV. For example, before DNA analysis, blood typing was extremely important for many criminal cases, not because of its PPV (same blood type between suspect and crime scene), but because of its NPV (different blood type between suspect and crime scene). Specifically, a negative match on blood type is definitive that the source was not from the suspect (exculpatory evidence), whereas a match is of little probative value, especially for common blood types. Both filler identifications and rejections from lineups have NPV as both behaviors decrease the likelihood that the suspect is the culprit (Wells, Yang, & Smalarz, 2015). Because ROC analysis on lineups ignores both of these eyewitness behaviors, we do not discuss NPV here, but interested readers are referred to Wells, Yang, and Smalarz (2015) for a thorough treatment of the diagnostic properties of different eyewitness behaviors.

Summarizing Performance in Identification Procedures and Deriving Theory

One of the most fundamental questions for researchers in any discipline is how to measure performance. Researchers examining eyewitness lineup procedures are no exception to this rule. Championing one identification procedure over another can be a difficult task. Often, procedures that decrease innocent suspect identifications also decrease culprit identifications (e.g., Clark, 2012; Cutler, 2013). That is to say that lineup procedures that benefit by decreasing innocent suspect identifications are also likely to come with a cost of decreasing culprit identifications. Accordingly, summary statistics are often used for the purpose of holistically comparing identification procedures. Traditionally, eyewitness researchers have relied on a family of ratios and proportions derived from Bayes's theorem for the purpose of comparing identification procedures. These Bayesian measures—sometimes called measures of probative value—reflect the trustworthiness of an identification decision and represent this trustworthiness in the form of likelihood ratios (diagnosticity ratio; Wells & Lindsay, 1980) or conditional probabilities (innocence risk; Clark & Godfrey, 2009; proportion guilty, Pryke et al., 2004). In recent years, however, some researchers have argued that eyewitness psychologists should abandon Bayesian measures in favor of ROC analysis or a theoretical proxy for the empirical ROC, the discriminability index (d'; Mickes, Moreland, Clark, & Wixted, 2014; Wixted & Mickes, 2012).

As noted above, the argument to abandon Bayesian measures in favor of ROC analysis rests on the assumptions (a) that ROC analysis of lineups measures underlying discriminability, and (b) that the procedure that produces superior underlying discriminability produces superior PPV. Based on these same assumptions, Wixted and Mickes (2014) derived a Diagnostic-Feature Detection theory of eyewitness identification to explain several patterns in ROC comparisons of identification procedures. According to Diagnostic-Feature Detection theory, when a suspect is presented to an eyewitness among a group of lineup members, it is apparent that some features are diagnostic (e.g., hairline, eye color) and others are not diagnostic (e.g., race, age), so eyewitnesses give little weight to nondiagnostic features. However, when a suspect is presented to an eyewitness in isolation, it never becomes apparent that some features are nondiagnostic and eyewitnesses do not discount the utility of these features. As the theory goes, discriminability is inhibited to the extent that eyewitnesses give weight to nondiagnostic features, so presenting the suspect to the eyewitness in a group of lineup members results in superior discriminability compared with when a suspect is presented to the eyewitness in isolation.

Diagnostic-Feature Detection is clearly a theory of psychological processes underlying discriminability. However, based on the description of the theory, it is not clear to us how the detection of diagnostic features increases underlying discriminability. Does the detection of diagnostic features increase culprit identifications? Maybe the detection of diagnostic features reduces mistaken identifications? We are not sure what the answer is. We note, however, that Diagnostic-Feature Detection has been proposed as an explanation for why fair lineups are superior to biased lineups and why fair lineups are superior to showups (Wetmore et al., 2015; Wixted & Mickes, 2014). And, because fair lineups do not yield a higher

rate of accurate identifications than either showups or biased lineups, then Diagnostic-Feature Detection theory must be a theory about mistaken identifications. However, as we will show, it is also the case that fair lineups do not produce fewer mistaken identifications than do biased lineups or showups. The key to understanding this conundrum is to recognize that ROC analysis, which is the analysis that spawned Diagnostic-Feature Detection theory, ignores a large proportion of the false positives that occur in lineups.

In our view, one must first account for how statistical dependencies in the data might explain differences in culprit and innocent suspect identifications before deriving psychological processing theories. Because the rate with which suspects are identified is dependent on the rate with which fillers are identified, one cannot simply set aside filler identifications and derive psychological processing theories based only on suspect identifications. To some extent, this reminds us of why researchers must use multilevel modeling when examining nested data. When data are nested, the classic regression assumption that error terms are independent is likely violated. Ignoring the dependencies among error terms is problematic because it can lead to spurious conclusions, so we use multilevel models to account for these statistical dependencies. Likewise, because suspect identification rates are dependent on filler identification rates, ignoring filler identifications and focusing only on suspect identifications can lead to the derivation of spurious processing theories. Making matters worse in the lineup context, suspect identifications are not the only response category indicative of memory performance; filler identifications are also indicative of memory performance. For these reasons, one cannot evaluate memory performance without considering both suspect and filler identifications and any psychological processing theory that attempts to explain memory performance in lineups must account for both suspect and filler identifications.

SDT and ROC Analysis in the Context of Lineup Research

SDT (e.g., Green & Swets, 1966; MacMillan & Creelman, 2005) is commonly used in basic memory research because the associated measures partition discriminability and response bias. Discriminability refers to the ability to distinguish between the presence and absence of stimuli (noise alone vs. signal + noise). Discriminability increases as signal strength increases (as separation between signal + noise and noise-only distributions increases) or as noise (i.e., variance) decreases (as signal + noise and noise-only distributions become more efficient). Response bias refers to the tendency (bias) to favor one response over some other response. In the context of eyewitness research, discriminability is the ability to distinguish between the culprit and innocent persons and response bias is the tendency to choose a lineup member as opposed to rejecting the lineup. In theory, response bias and discriminability should reflect both suspect and filler choices, but the ROC approach advocated by Wixted and Mickes (2012) lumps filler identifications and rejections into a single nonidentification category. This is problematic as this practice makes response bias appear more conservative than it really is, obscures our understanding of discriminability, and hides important and regularly occurring phenomena in lineup procedures (e.g., Wells, Smalarz, & Smith, 2015; Wells, Smith, & Smalarz, 2015).

ROC curves were developed for problems involving binary classification (Egan, 1958). Binary classification problems are those for which there are two potential states of the world (noise only, signal + noise) and two potential responses by the receiver (negative response, affirmative response). The ROC curve is then graphed by plotting the true-positive rate (on the ordinate) against the false-positive rate (on the abscissa) over a series of decision criteria. However, eyewitness lineups do not fit the traditional 2 (noise only, signal + noise) × 2 (negative response, affirmative response) confusion matrix for which ROC analysis was designed. Indeed, lineups have a 3 (innocent suspect, filler, culprit) × 2 (negative response, affirmative response) confusion matrix that needed to be repackaged before it could be subjected to ROC analysis. Wixted and Mickes (2012) proposed a solution in which they plotted culprit identifications on the ordinate against innocent suspect identifications on the abscissa and this is the method that researchers using ROC analysis on lineup data have followed ever since. In effect, this ROC approach means that false-positive filler identifications and rejections are treated as if they were the same thing for both theoretical and applied purposes.

After an eyewitness makes an identification decision from a lineup, researchers routinely ask the eyewitness to express his or her level of confidence in the decision. In plotting ROC curves for lineup procedures, the expressed level of eyewitness confidence is used to estimate decision criterion. Each eyewitness confidence level represents a different decision criterion, so we use the term decision criteria throughout the manuscript to reflect the fact that ROC curves are characterized by multiple decision criteria. Graphing ROC curves for identification procedures works like this: the leftmost point is first plotted in the ROC space. This point includes the proportion of culprit (reflected on the ordinate) and innocent suspect identifications (reflected on the abscissa) that occur at only the highest level of eyewitness confidence. The next point reflects the proportion of culprit and innocent suspect identifications occurring at the highest and second to highest levels of confidence. This procedure continues until the rates of culprit and innocent suspect identifications collapsed over all confidence levels is reflected in a single point (the rightmost point of the ROC curve). After all of the points have been plotted in the ROC space, a curve is drawn through these points extending from the lower left corner of the ROC space (0,0) to the rightmost point of the ROC. Unlike traditional Old/New facial recognition tasks, participants in eyewitness-lineup experiments generally only make one recognition judgment and a single ROC curve is generated for each between-subjects condition. After the ROC curve is plotted, the area under the ROC curve (AUC) is calculated in an attempt to estimate the discriminability associated with a given identification procedure. The AUC value reflects the total area of the ROC space falling below the ROC curve.

Because lineups include both a suspect and fillers, the innocent suspect identification rate is unlikely to ever reach a proportion of 1.00 as some false-affirmative eyewitness responses will land on fillers instead of the innocent suspect. Indeed, in a perfectly fair six-person lineup (a lineup in which all fillers are as likely to be identified as the innocent suspect), the maximum innocent suspect identification rate could not exceed .167 (1/6). Given that ROC comparisons of identification procedures only plot false-affirmative identifications that land on the innocent suspect (and not those that land on fillers) the ROC curve is restricted to the far

left of the ROC space and researchers only examine a partial AUC (pAUC). The pAUC extends from an abscissa value of 0 (an innocent suspect identification rate of 0) to some other value less than 1. One can compare identification procedures by plotting their respective curves in the same ROC space and the procedure that produces the larger area under the curve is argued to have superior discriminability.

It is important to note from the outset that we believe that ROC analysis is perfectly appropriate for studies employing the 2 (noise, signal + noise) × 2 (affirmative response, negative response) confusion matrix for which ROC analysis was originally designed. We do not have a problem with ROC analysis per se, and we believe it is a useful tool when used on problems that fit the standard 2 × 2 confusion matrix structure. However, we do have a problem with lineup researchers forcing the 3 (innocent suspect, filler, culprit) × 2 (affirmative response, negative response) lineup structure into the 2 × 2 structure to conduct an ROC analysis and claiming that observed differences reflect differences in underlying discriminability. This repackaging of the lineup matrix is achieved by disguising false alarms as misses (false rejections) in culprit-present lineups and as correct rejections in culprit-absent lineups. When researchers violate the long held assumption of SDT that false-affirmative responses are false-affirmative responses and not rejections, we argue that ROC analysis does not measure underlying discriminability and can lead to erroneous conclusions surrounding the theoretical and applied utility of lineups.

ROC Analysis of Lineups Does Not Measure Underlying Discriminability and Misguides Theory

In this section we address the first assumption made by those who advocate using ROC analysis to compare lineup procedures, namely that ROC analysis of lineups measures underlying discriminability. We first point out that it has long been recognized by Signal Detection theoreticians that ROC analysis does not measure underlying discriminability when tasks exceed the dimensions of the 2 (noise, signal + noise) × 2 (negative response, affirmative response) confusion matrix and more sophisticated analytical tools have been developed to measure discriminability in these situations (Duncan, 2006; Starr, Metz, Lusted, & Goodenough, 1975). We then demonstrate how the mistaken assumption that ROC analysis does measure underlying discriminability in this circumstance led to the development of a processing theory with no empirical bases, namely, Diagnostic-Feature Detection (Wixted & Mickes, 2014).

False Alarms Are Never Rejections

When using ROC analysis to compare lineup tasks, researchers treat filler identifications as rejections. The logic for this decision is that, unlike an innocent suspect, if a filler is identified, this is not a dangerous error, because the filler is a known innocent and will not be prosecuted. This is certainly true, but it is also completely irrelevant if the goal of ROC analysis is to assess underlying discriminability. The selling point of ROC analysis for lineups is that it can get at underlying discriminability, and the assumption that ROC analysis on lineups assesses underlying discriminability is what led Wixted and Mickes (2014) to propose Diagnostic-Feature Detection as a psychological processing theory. It has long been recognized in other applied fields such as radiology—which Mickes, Flowe, and Wixted (2012) liken to an eyewitness lineup—that if the goal of a procedure is to get at underlying discriminability, then false alarms must be treated as false alarms and not as rejections (Starr et al., 1975; see also Macmillan & Creelman, 2005, pp. 256–258) and more sophisticated models for dealing with compound decisions have been proposed in recent years (e.g., Duncan, 2006; Palmer & Brewer, 2012).

Let us consider the radiology study by Starr et al. (1975) more closely, because, unlike most X-ray tasks, the task used in Starr et al. (1975) actually was similar to a lineup task. Like a lineup, the X-ray task used by Starr et al. (1975) was a compound-decision task involving both detection and identification. Participants were presented with X-rays divided into four quadrants and the signal (a tumor) was either present or absent. When the signal was present the task of the participant was to (a) correctly *detect* the presence of the signal and (b) correctly *identify* which of the four quadrants the signal was present in. When the signal was absent (noise-only trials), the task of the participant was to indicate that the signal was not present. From these data, the researchers plotted a Location-ROC (LROC).[1] Like a lineup task, a hit only occurs on trials in which participants both detect the presence of the signal and correctly identify the quadrant the signal was present in; however, anytime participants detect the presence of the signal when it is not present, the participant has committed a false alarm (it does not matter which quadrant the participant chose). The practice of treating culprit-absent filler identifications from lineups as correct rejections is in direct contrast with the work of Starr et al. (1975).

Although LROC analysis correctly classifies filler identifications in the culprit-absent lineup as false alarms, it still misclassifies filler identifications in the culprit-present lineup only as misses (when they are simultaneously both a miss and a false alarm). Because of the misclassification of filler identifications as only misses in the culprit-present lineup, LROC analysis cannot perfectly separate discriminability and response bias, but it certainly appears to be a step in the right direction and a more suitable analytical framework than ROC analysis if the goal is to measure the underlying discriminability associated with lineups.

Another method for examining the underlying discriminability associated with compound-decision tasks is the SDT-compound decisions model (SDT-CD; Duncan, 2006; Palmer & Brewer, 2012). Because the focus of SDT-CD is on decision-making performance, all identification decisions are considered in this model. Like the LROC, SDT-CD has two components, (a) detection (i.e., is the culprit present?) and (b) identification (i.e., which stimulus is the culprit?). Also like the LROC, the identification of any lineup member in a culprit-absent lineup (filler or innocent suspect) is a false alarm. Where the SDT-CD model offers improvement over the LROC approach is in its ability to simultaneously classify target-present filler identifications as both misses and false alarms. This is extremely important if the goal is to measure

[1] Starr et al. (1975) referred to this as an Identification Operating Charcteristic (IOC), but it has been referred to as an LROC elsewhere in the eyewitness literature (e.g., Wixted & Mickes, 2015a), so we adopt that terminology here for consistency.

underlying discriminability. Consider the two hypothetical procedures described by Palmer and Brewer (2012). Both have identical culprit identification rates of 40% and false-alarm rates (i.e., choosing of any culprit-absent lineup member) of 20%. However, procedure A has a culprit-present filler identification rate of 60% and procedure B has a culprit-present filler identification rate of 0%. Which procedure is more impressive? Clearly procedure B is more impressive. Many of the culprit identifications in procedure A may not be because of the ability of the eyewitness to discriminate between old and new faces, but to a liberal response bias. SDT-CD offers a significant improvement over ROC or LROC analyses as it accounts for culprit-present filler identifications and appears to better distinguish between response bias and discriminability.

In contrast to earlier work, SDT-CD assumes that a single discriminability parameter guides performance on both the detection and identification components of compound-decision tasks. The detection component measures how well eyewitnesses can detect the presence of the culprit in a lineup, regardless of whether the eyewitness can determine which lineup member is the culprit. Perfect detection performance would occur under the condition that every eyewitness who was presented with a culprit-present lineup made an identification (regardless of whether they chose the culprit or a filler) and every eyewitness who was presented with a culprit-absent lineup rejected the lineup. The identification component measures how well eyewitnesses can sort between the culprit and fillers. Keeping the identification component of SDT-CD distinct from the detection component, identification performance reflects the conditional probability that an eyewitness will identify the culprit given that the eyewitness chose someone from the culprit-present lineup. Perfect identification performance occurs under the condition that 100% of identifications from the culprit-present lineup are of the culprit (Duncan, 2006; Palmer, Brewer, & Weber, 2010; Palmer & Brewer, 2012).

Because there is no analytical solution for determining the single discriminability and response bias parameters that underlie performance in compound-decision tasks, SDT-CD employs a model-fitting approach. Specifically, SDT-CD involves finding the single discriminability and response bias parameters that minimize the discrepancy between model-expected eyewitness performance and observed eyewitness performance. See Duncan (2006) for a more thorough treatment on SDT-CD.

Both SDT-CD and LROC analysis offer significant improvements over ROC analysis if the goal is to measure underlying discriminability. Because ROC analysis misclassifies culprit-present filler identifications as misses and culprit-absent filler identifications as correct rejections, it dramatically underestimates response bias and overestimates discriminability. Indeed, by classifying culprit-absent filler identifications as correct rejections, ROC analysis of lineups overestimates the proportion of correct decisions made by eyewitnesses. LROC better estimates underlying discriminability than does ROC analysis because all false-affirmative responses in culprit-absent lineups are correctly classified as false-affirmative responses. LROC may still slightly overestimate discriminability, however, as false-affirmative filler identifications in culprit-present lineups are classified as misses and not factored into calculations of response bias. Finally, SDT-CD appropriately classifies all decisions from the 3 (rejection, filler identification, suspect identification) $\times$ 2 (culprit present, culprit absent) lineup matrix. The one slight drawback to using SDT-CD to measure decision-making performance in lineups is that because there is no analytical solution, one must rely on a model-fitting approach to determine the best-fitting decision parameters.

There are many theoretical questions surrounding lineups for which one might want to know which procedure facilitates superior underlying discriminability. If the goal of an analysis is to measure underlying discriminability from lineups, LROC analysis or SDT-CD are more suitable alternatives than ROC analysis insofar as they take into account all the judgments that emerge from lineup procedures. Knowing which of two identification procedures produces better underlying discriminability is important for theoretical purposes. However, as we will demonstrate below, superior underlying discriminability does not necessitate superior PPV. Accordingly, it does not follow that the procedure that produces a superior LROC or d'' is necessarily the procedure that produces superior forensic outcomes.[2] In the remainder of this section we demonstrate how ROC analysis on lineups led Wixted and Mickes (2014) to make unwarranted inferences about underlying discriminability and to develop a processing theory for which there is no empirical support.

An Empirical Test of Diagnostic-Feature Detection Theory: Lineups Versus Showups

Showups are a one-person identification task in which a single suspect who is guilty or innocent is presented to the eyewitness for an identification test. Unlike lineups, showups do not include fillers. In some sense, showups resemble the traditional Old/New facial recognition task.[3] The comparison of fair, simultaneous lineups and showups offers a strong initial test of Diagnostic-Feature Detection theory. Because showups include only a single person (the suspect) they allow for a direct test of the extent to which surrounding the suspect with other persons (as is done in a simultaneous lineup) facilitates the ability of the eyewitness to discriminate between the culprit and innocent people in the lineup. According to Wixted and Mickes (2014), lineups are superior to showups because:

> When a face is presented in isolation . . . there is no obvious indication to the eyewitness that some features are diagnostic and others are not. To the extent that the non-diagnostic features are given weight under those circumstances, the ability to discriminate innocent from guilty suspects will suffer. In a simultaneous lineup . . . it is immediately

[2] Consistent with past work (Palmer & Brewer, 2012), we refer to the discriminability index from SDT-CD as d'' to distinguish it from the more common simple-decision statistic d'. We also refer to d' values calculated without including filler identifications as PPVd' to distinguish such values from measures of underlying discriminability, that is, d' for 2 $\times$ 2 tasks and d'' for compound-decision tasks.

[3] Because showups are used in different contexts than lineups (or Old/New recognition tasks), there are differences that make showups unique. For instance, they are often used in the field and are host to an array of social influences and pressures (e.g., the presence of property stolen during the crime, wearing the same clothing the culprit was in, hearing over police radio that the police have the culprit, etc.) that might not be present in lineup tasks and certainly are not present in facial recognition tasks. However, if one were to strip away these social influences and pressures from a showup task, there would be little to distinguish it from a standard Old/New recognition task.

> apparent to the eyewitness that everyone in the lineup shares certain non-diagnostic features. . . . For that reason, the eyewitness will be encouraged to attach weight to features that might be diagnostic while discounting features that are non-diagnostic. (p. 269)

Diagnostic-Feature Detection, then, is a theory about how to improve underlying discriminability by surrounding a suspect with good fillers (i.e., fillers who resemble the culprit to the same extent as the innocent suspect). Moreover, Diagnostic-Feature Detection predicts that eyewitnesses would be better able to reject a culprit-absent lineup than they would be able to reject a culprit-absent showup. This is evident from Diagnostic-Feature Detection theory, because, to the extent that eyewitnesses presented with showups give more weight to nondiagnostic features than they do with a lineup, witnesses should be less likely to make an identification from a culprit-absent lineup than from a culprit-absent showup.

In a test of this proposition, Wetmore et al. (2015) compared simultaneous lineups and showups using ROC analysis and attributed the superiority of lineups to Diagnostic-Feature Detection theory. Wells et al. (2015b, 2015c) recently reanalyzed these data to demonstrate that the evidence actually contradicts that theory. Particularly important is the fact that witnesses are actually more likely, not less likely, to make a false-affirmative identification from a culprit-absent lineup than from a culprit-absent showup. However, as has been long documented in the eyewitness literature, these false-affirmative identifications spread across the members of the absent lineup and thereby reduce the rate of false identifications that land on the innocent suspect (Wells, 2001). More specifically, the data show a different process that accounts for the superiority of lineups over showups, a process called differential filler siphoning. Differential filler siphoning is a process through which fillers draw false-positive responses away from the innocent suspect to a greater extent than they draw positive responses away from the culprit. Differential filler siphoning is—as we demonstrate below—predicted by SDT. What is at dispute here is not that lineups are superior to showups at the applied level; that was established long ago (e.g., Steblay, Dysart, Fulero, & Lindsay, 2003). What we dispute here is *why* lineups are superior to showups. To explore this issue, we now consider Wells et al.'s (2015b, 2015c) reanalysis of Wetmore et al.'s (2015) study comparing simultaneous lineups and showups.

Consider the data in Table 1, which were originally presented in Wells et al. (2015b) based on data published by Wetmore et al. (2015). In Panel A of Table 1, it is evident that 64.4% of eyewitnesses made a false-positive error from the lineup (10.2% innocent suspect identifications, 54.2% filler identifications); however, the way these data are repackaged from the natural 3 × 2 into a 2 × 2 for ROC analysis (Panel B of Table 1) would lead one to believe that only 10.2% of eyewitnesses made a false-positive error. If one focuses on the data from Wetmore et al. (2015) as they repackaged it for ROC analysis (Panel B), one would likely conclude that discriminability was good for the lineup because 68.3% of eyewitnesses identified the culprit when present and 89.8% of eyewitnesses correctly rejected when the culprit was absent. However, we can clearly see when examining the entire 3 × 2 (Panel A of Table 1) that this is not what happened with the lineup; instead, only 35.6% made correct rejections. Here, we see a serious misreading of the data from the ROC approach (Panel B) because of treating filler identifications in the absent lineup as if they were correct rejections.

By comparing the ROC representation of lineups (Panel B) to showups (Panel C, Table 1), Wetmore et al. (2015) incorrectly concluded that the reason lineups were superior to showups was because they increased discriminability. Wetmore et al. (2015) went on to interpret this effect in terms of Diagnostic-Feature Detection, by claiming that simultaneous lineups increased discriminability relative to showups

> . . . because multiple lineup members can be compared. This allows diagnostic and non-diagnostic features to be distinguished, and the diagnostic features to subsequently receive more attention. A showup

Table 1
Lineup and Showup Data From Wetmore et al. (2015; Fair/Immediate Conditions)

	ID suspect	ID filler	Rejections
Panel A. The Actual 3 × 2 Lineup Data From Wetmore et al.			
Culprit present	68.3% (true positives; accurate identifications of culprit)	10.0% (false positives; mistaken IDs of innocent fillers)	21.7% (false rejections)
Culprit absent	10.2% (false positives; mistaken IDs of the innocent suspect)	54.2% (false positives; mistaken IDs of innocent fillers)	35.6% (correct rejections)

Panel B. Lineup Data From Wetmore et al. After Being Forced Into a 2 × 2 for Purposes of ROC Analysis

	ID suspect	"Rejection"
Culprit present	68.3%	31.7% (actually, this 31.7% includes both false rejections and false positive IDs of fillers; see Panel A)
Culprit absent	10.2%	89.8% (actually, this 89.8% includes both correct rejections and false positive IDs of fillers; see Panel A)

Panel C. Showup Data From Wetmore et al.

	ID suspect	Rejection
Culprit present	62.1% (true positives; accurate identifications of culprit)	37.9% (false rejections)
Culprit absent	42.0% (false positives; mistaken IDs of the innocent suspect)	58.0% (correct rejections)

Note. ROC = Receiver Operating Characteristic.

> does not allow this comparison, and consequently, diagnostic features may never become apparent. (p. 13)

Lineups are superior to showups, but not because they increase discriminability. A careful examination of Panel A (see Table 1) reveals that only 35.6% of eyewitnesses correctly rejected the culprit-absent lineup, not 89.8% as the ROC representation would lead one to believe. The fact that only 10.2% of eyewitnesses identified the innocent suspect is irrelevant to the discussion of underlying discriminability. The lineup did not decrease innocent suspect identifications by increasing correct rejections relative to the showup. In fact, the showup procedure produced more correct rejections (58%) than did the lineup (35.6%). If the lineup decreased innocent suspect identifications, but did not increase correct rejections, then how was this accomplished? The lineup accomplished this by adding fillers to the identification procedure. In the absence of fillers, the innocent suspect in the showup was forced to shoulder all of the false alarms (42%) himself. But, in the lineup, the 64.4% false-alarm rate was distributed among lineup members and the innocent suspect was only forced to shoulder 10.2% of those false alarms. This is not evidence that lineups increase discriminability. This is evidence that lineups redistribute false alarms, a phenomenon referred to as differential filler siphoning. Differential filler siphoning refers to the idea that good fillers siphon more choices from the innocent suspect than they do from the culprit. Why is the siphoning differential (i.e., more siphoning from the innocent suspect than from the culprit)? The reason that good fillers siphon more from the innocent suspect than from the guilty suspect is because the good fillers are as similar (on average) to the witnesses' memories of the culprit as the innocent suspect (hence, they compete effectively for choices) but the fillers are not as similar to the witnesses' memories of the culprit as the culprit himself is (hence, tend not to compete as effectively for choices when the culprit is present).

Why did the fair lineup produce a slightly higher culprit identification rate than did the showup? That slight difference in culprit identification rates is nonsignificant, as it typically is in lineup versus showup studies (Clark, 2012; Steblay et al., 2003). However, it is important to note that lineups can produce a slight increase in culprit identifications to the extent that they exact a more liberal decision criterion than showups. And, in fact, Wetmore et al.'s (2015) data show that response bias was more liberal in the culprit-present lineup than the culprit-present showup (16.2% more choosing in the culprit-present lineup than the culprit-present showup) and the majority of these additional choices (10.0% of that 16.2%) in the culprit-present lineup landed on fillers. Thus, this bump in culprit identifications for the lineup is, at least in part, attributable to the more liberal response bias associated with the lineup procedure.

A Second Empirical Test of Diagnostic-Feature Detection Theory: Fair Versus Biased Lineups

Not only does Diagnostic-Feature Detection predict that discriminability will increase as the number of fillers in a lineup increases, but it also predicts that discriminability will increase as the similarity between fillers and the culprit increases. A result of this prediction is that Diagnostic-Feature Detection predicts fair lineups to have superior discriminability compared with biased lineups. A lineup is fair to the extent that an innocent suspect is only identified by chance among those who identify someone. For example, in a six-person culprit-absent lineup only 16.67% (1/6) of lineup choices should land on the innocent suspect. To the extent that more than 16.67% of lineup choices land on the innocent suspect, the lineup is biased toward the innocent suspect. In some sense, a showup can also be thought of as a biased lineup. Because showups only include a suspect and no fillers, an innocent suspect would always stand out and would be forced to shoulder all false-affirmative responses.

In addition to fair lineups and showups, Wetmore et al. (2015) also had a biased lineup. In the culprit-absent condition, the innocent suspect was identified 28.1% of the time and each of the fillers was only identified, on average, 6.32% of the time. The innocent suspect stood out from the other members of the lineup—the defining feature of a biased lineup. The fillers did not offer as much competition for eyewitness choices (fillers were less familiar than the innocent suspect) as did those fillers in the fair lineup (innocent suspect identification rate: 10.2%, average filler identification rate: 10.8%). In the biased culprit-present lineup, the culprit was identified 81.4% of the time and each of the fillers was only identified, on average, 0.58% of the time. Biased-lineup fillers were less able to compete with eyewitness choices in the biased culprit-present lineup than were those fillers in the fair culprit-present lineup (culprit identification rate: 68.3%, average filler identification rate: 2%). However, the benefit of the biased lineup (a 13.2% increase in culprit identifications) was less than the cost of the biased lineup (a 17.3% increase in innocent suspect identifications) as is reflected by the fact that the culprit was only 2.90 times more likely to be identified than was the innocent suspect in the biased lineup (vs. 6.69 times in the fair lineup).

Moreover, the superiority of fair lineups in terms of PPV exists despite the fact that biased lineups are actually associated with better discriminability. In the fair lineup, 68.3% of eyewitness correctly identified the culprit when present and only 35.6% of eyewitnesses correctly rejected the lineup when the culprit was absent. In the biased lineup, 81.3% of eyewitnesses identified the culprit when present and 40.3% of eyewitness correctly rejected the lineup when the culprit was absent. Hence, biased lineups produced 13% more culprit identifications *and* 4.7% more correct rejections than the fair lineups. Given that biased lineups produced more culprit identifications and more correct rejections, it is objectively clear that biased lineups produced superior underlying discriminability. But, to further test this proposition, we fit SDT-CD models to the fair and biased lineups presented in Wetmore et al. (2015). SDT-CD provided adequate fits to both the fair and biased lineup procedures, G^2s(3) < 3.30, ps $> .35$.[4] As predicted, the best-fitting decision parameters for the biased-lineup were associated with superior underlying discriminability ($d'' = 2.55$) when compared with the best-fitting parameters for the fair-lineup procedure ($d'' = 1.91$). Moreover, the best-fitting model parameters generated by SDT-CD showed little evidence of a shift in response bias, and if anything, response bias was more conser-

[4] The G^2 statistic is conceptually the same as a χ^2 test, but G^2 statistics can be summed together. Accordingly, we calculated G^2 statistic for target-absent filler identifications, target-present filler identifications, and culprit identifications. We then summed these statistics together to assess model fit. See Palmer and Brewer (2012).

vative for the biased lineup ($c = -.19$) than for the fair lineup ($c = -.28$). This is also inconsistent with the ROC narrative as researchers often interpret ROC analysis on lineups as reflecting response bias and biased lineups result in more suspect identifications than do fair lineups. But, as is evident from considering the full 3 × 2 and as is confirmed by SDT-CD, eyewitnesses make more total identifications (suspect plus fillers) from fair lineups and thus, fair lineups are associated with a more liberal response bias. The predicted response probabilities from the SDT-CD model are presented in Table 2.

The fact that biased lineups have superior underlying discriminability when compared with fair lineups directly contradicts any argument that ROC analysis on lineups measures underlying discriminability. Wetmore et al.'s (2015) original ROC analysis found that the fair lineup produced a higher ROC curve than did the biased lineup. Thus, ROC analysis on lineups cannot possibly be a measure of underlying discriminability otherwise the biased lineup would have a higher ROC than the fair lineup. What ROC analysis on lineups does appear to reflect is PPV, as is evidenced by the fact that the fair lineup ROC was higher than both the biased lineup and showup ROCs. In the final section of the manuscript we will introduce a formal model that predicts this dissociation between PPV and underlying discriminability when comparing fair lineup procedures with either biased lineup procedures or showups. It should also be noted at this point that this pattern of results directly contradicts the predictions of Diagnostic-Feature Detection theory. Based on the assumption that ROC analysis on lineups measures underlying discriminability and the finding that fair lineups result in higher ROC curves than biased lineups, Diagnostic-Feature Detection was offered as an explanation as to why fair lineups increase underlying discriminability when compared with biased lineups (Wixted & Mickes, 2014, 2015a, 2015b). It is now clear that the Diagnostic-Feature Detection hypothesis can be rejected given that biased lineups have better, not worse, discriminability than do fair lineups.

Computational Evidence for Differential Filler Siphoning

Wixted and Mickes (2015b) recently argued that differential filler siphoning could not explain why fair lineups result in higher ROC curves than showups or biased lineups. Their argument rested on two demonstrations. The problem with these demonstrations is that neither demonstration measures underlying discriminability.

In their first demonstration, Wixted and Mickes (2015b) fit a SDT model to the Wetmore et al. (2015) data and the resulting d' on suspect identifications only was much larger for the lineup (1.63) than the showup (0.70). Henceforth, we refer to d' on suspect identifications as PPVd' to reflect the fact that d' on suspect identifications only is not a measure of underlying discriminability, but is an applied measure of PPV. In their second demonstration, Wixted and Mickes (2015b) equated PPVd' and decision criteria for the two procedures, ran a simulation, and the expected values generated by the simulation provided a worse fit to the data than their first simulation in which they permitted PPVd' to vary between showups and lineups. Moreover, the two predicted ROCs from this second simulation fell atop one another until the lineup ROC approached its maximum false-alarm rate (.167) at which point the showup ROC was slightly higher. Because filler siphoning was occurring in the lineup and because PPVd' was the same in both procedures, they argued that filler siphoning could not explain why the lineup procedure produced a higher ROC curve in Wetmore et al.'s (2015) lineups, but Diagnostic-Feature Detection could (because if differential filler siphoning was occurring for lineups, then the ROC curves should reflect this). Wixted and Mickes (2015b) argued that this demonstrates that lineups produce superior underlying discriminability relative to showups and that "this result corresponds to what one would immediately infer by examining the objective ROC data . . . and contradicts the claim by Wells et al. (2015) that theoretical discriminability is not higher for lineups" (Wixted & Mickes, 2015b, p. 331).

In the first demonstration, Wixted and Mickes (2015b) found that PPVd' was higher for lineups than for showups. All this tells us is that the difference in culprit and innocent suspect identifications is greater for lineups than it is for showups, that is, that PPV is higher for lineups. However, that result was never in question and is readily apparent from examining the observed data. Simply calculating PPVd' for the observed lineups ($d' = 1.74$) and showups ($d' = 0.51$) reveals this same information. Nobody disagrees that lineups have superior PPV when compared with showups. But, this first demonstration proffered by Wixted and Mickes (2015b) does not tell us *why* lineups have superior PPV when compared to showups. Do lineups have superior PPV when compared with showups because they increase underlying discriminability, because of differential filler siphoning, or because of some other reason? This is the only question that is debated and this first demonstration proffered by Wixted and Mickes (2015b) does not address this question.

The second demonstration proffered by Wixted and Mickes (2015b) also fails to address the question of why lineups have superior PPV than do showups. Indeed, Wixted and Mickes (2015b) equated PPVd' for the two procedures and the fit was worse than in their first demonstration where they permitted PPVd' to vary between lineups and showups. All this tells us is that a model in which the difference in culprit and innocent suspect identifications is greater for the lineup than for the showup provides a better fit to the observed data than does a model in which the difference in culprit and innocent suspect identifications is the same for the lineup as it is for the showup. Of course this is the case, because we know from the observed data that the difference

Table 2
Observed and Model-Predicted Response Probabilities, Model Fit Statistics (G^2), and Estimates of Discriminability (d'') and Response Bias for Wetmore et al.'s (2015) Fair and Biased Lineups

	Observed			Model					
Lineup	CID	FID	FA	CID	FID	FA	G^2	d''	c
Fair	.68	.10	.64	.64	.22	.61	3.30	1.91	−.28
Bias	.81	.03	.60	.78	.11	.57	2.07	2.55	−.19

Note. CID = culprit identification rate; FID = culprit-present filler identification rate; FA = culprit-absent false alarm rate (filler identifications plus innocent suspect identifications).

between culprit and innocent suspect identifications is greater for the lineup than the showup. Like their first demonstration, the second Wixted and Mickes (2015b) demonstration also fails to address the critical question of why lineups have superior PPV than do showups.

Wixted and Mickes (2015b) seem to think that only the procedure with superior underlying discriminability can have a higher PPVd'. That is not true. As we have already shown, when compared to biased lineups and showups, fair lineups have inferior underlying discriminability; yet, fair lineups have superior PPV, that is, higher PPVd' and higher diagnosticity ratio on suspect identifications. Once one appreciates that PPVd', whether derived from observed or simulated data, is a measure of PPV and not underlying discriminability, it becomes obvious that the evidence Wixted and Mickes (2015b) offered for Diagnostic-Feature Detection theory actually provides compelling evidence for differential filler siphoning. Indeed, underlying discriminability in their first demonstration was better for the showup than the lineup (fewer false-positive errors in the showup, similar culprit identification rates in both procedures); yet, the lineup has superior PPV (PPVd' was much higher for the lineup than for the showup). This is precisely the pattern of results predicted by a model of differential filler siphoning.

Wixted and Mickes (2015b) did not appreciate that the only way to demonstrate increased PPV in their demonstrations was to increase PPVd'. Any process (e.g., differential filler siphoning, Diagnostic-Feature Detection, etc.) that increased PPV would lead to an increase in PPVd'. Accordingly, there is little that we can learn from their simulations if we focus only on PPVd'; however, once we consider that underlying discriminability was worse in the lineup procedure and yet, the lineup procedure had superior PPV (higher PPVd') than did the showup, it becomes apparent that their demonstrations provide fairly compelling evidence for the differential filler siphoning explanation.

Although PPVd' clearly does not measure underlying discriminability from lineup procedures, we discovered something equally disconcerting when examining the Wixted and Mickes (2015b) simulations. As it turns out, these simulations make the untenable assumption that every eyewitness constructs exactly the same decision criteria. But, even the earliest works on SDT recognized that different subjects would have different criteria for making an affirmative response (e.g., Green & Swets, 1966). Moreover, many researchers now argue that decision criteria might even vary *within* a given participant (e.g., Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008; cf. Kellen, Klauer, & Singmann, 2012, 2013). In other words, just as signal strength has a probabilistic distribution, so do decision criteria, even for within-subjects designs. Furthermore, the assumption that every witness has exactly the same decision criteria are especially likely to be false in a between-subjects design in which a given subject contributes only one data point (a characteristic of most eyewitness identification experiments).

Benjamin et al. (2009) call this variation in decision criteria *criterial noise*. We will call it *criterial variance* to reflect the idea that it represents variance from one eyewitness to another in their criteria for making an identification decision. Indeed, because decision criteria are subjective, there seems little room for debate that there are individual differences in decision criteria and that this should be represented in any credible computational simulation of eyewitness identification processes. Therefore, we did a computational simulation that replicated what Wixted and Mickes (2015b) found and then did that same simulation but allowing for criterial variance. As we will show, the results are very different when a more realistic model is used that includes criterial variance. As a matter of fact, once criterial variance is permitted, it is no longer necessary to represent the benefits of differential filler siphoning by increasing PPVd'. When criterial variance is permitted, PPVd' can be held constant and the increased PPV that comes from differential filler siphoning still emerges.

Computational Models Reveal the Benefits of Differential Filler Siphoning but Only When Criterial Variance Is Allowed

We began by running a simulation of lineups versus showups using methods very similar to Wixted and Mickes' simulation (2015b; also see Lampinen, 2016) so as to replicate their finding that ROC curves for lineups were not superior to showups using such a simulation. Like Wixted and Mickes (2015b), we set PPVd', decision criteria, and other relevant parameters to be the same for lineups and showups (see Appendix A for a full description of the parameter settings). The result is shown in Figure 1, which replicates what Wixted and Mickes found; the lineup and showup ROC curves fall atop one another until the lineup reaches its maximum innocent suspect identification rate (.167) at which point there is a slight showup advantage. We then ran the same simulation except that we used the more realistic assumption that not every witness uses the same exact decision criteria and instead assumed that there is criterial variance (see Appendix A for the values of criterial variance). The result of this simulation is shown in Figure 2. As can be seen in Figure 2, when criterial variance is permitted, the lineup produces an ROC curve that is clearly higher than that of the showup. The higher ROC curve for lineups than for showups in Figure 2 occurs despite the fact that d' on suspect identifications, mean decision criteria, and all other relevant parameters were equal for the lineup and the showup.

Figure 2 shows clearly that as long as criterial variance is permitted, differential filler siphoning alone can produce a higher ROC curve for lineups than for showups. There is no need to presume that the presence of fillers somehow improves discriminability (through Diagnostic-Feature Detection or any related mechanism) to obtain a higher ROC curve for lineups than for showups. Differential filler siphoning, in which fillers draw more choices away from the innocent suspect than they draw from the guilty suspect, explains this phenomenon. We cannot think of any subjective human judgment that does not have between-subjects variance, especially something like subjective decision criteria, so we believe that the assumption that every witness has the same exact decision criteria are untenable and that the simulation in Figure 2 is more realistic.

We note that although the ROC curve is higher for the lineup than for the showup, discriminability was actually lower for the lineup than for the showup. This is readily evident in the fact that total false alarms plus false rejections were higher for the lineup than for the showup whereas total hits plus correct rejections were higher for the showup than for the lineup. That cannot happen if lineups have higher discriminability than showups. The ROC curves are higher for the lineup than for the showup because ROC

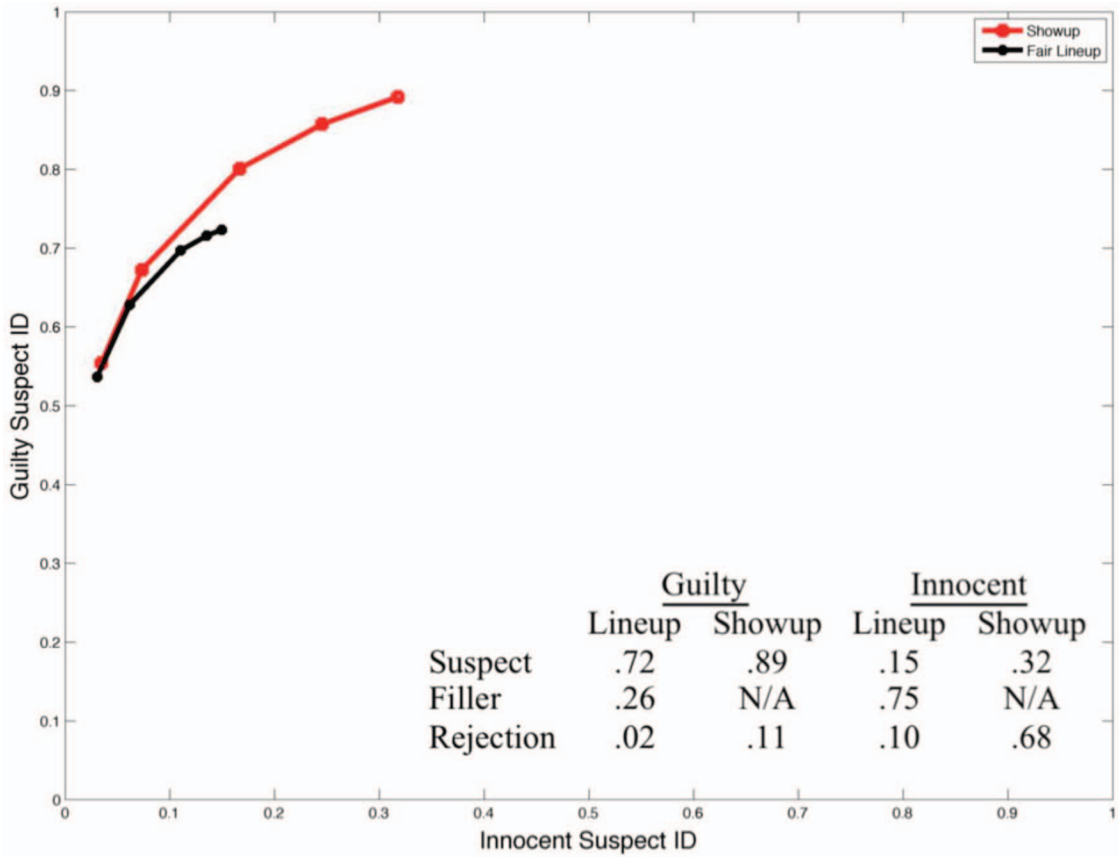


Figure 1. Computational simulation comparing showups (red square markers) to fair lineups (black circular markers) when it is assumed that every witness uses exactly the same decision criterion. See the online article for the color version of this figure.

curves for lineups are based on only two of the six cells in the 3 × 2 lineup structure, thereby reflecting PPV, and fail to reflect underlying discriminability. What we have shown here is the dissociation between ROC curves and underlying discriminability when ROC curves are used with lineups.

Computational simulation of fair versus biased lineups. We end this section with one more computational simulation to show how differential filler siphoning can result in a higher ROC curve without any change in actual d' on suspect identifications for fair versus biased lineups. According to Wixted and Mickes' (2015a) Diagnostic-Feature Detection theory, good fillers help the witness to decide which features are relevant and which features are irrelevant so as to make better decisions (i.e., increase discriminability via a comparison process), whereas poor fillers do not serve this improved discriminability function. Wells et al. (2015b, 2015c), on the other hand, argue that discriminability is not better with fair than with biased lineups. Instead, fair lineups result in better PPV only because good fillers result in differential siphoning and poor fillers result in little or no siphoning and the benefit of differential filler siphoning is so great that it is able to overcome inferior discriminability in the fair lineup. Both the Diagnostic-Feature Detection account and the differential filler siphoning account predict that ROC curves will be higher for fair than for biased lineups.

If increased discriminability is required for the higher ROC in a fair versus biased lineup, however, then a computational simulation that allows only for differential filler siphoning and not for some additional mechanism that changes discriminability, should not show a higher ROC curve for fair lineups than for biased lineups.

Hence, we ran another simulation, this time for fair versus biased lineups. We set d' on suspect identifications, decision criterion, and other relevant parameters to be the same for fair and biased lineups (see Appendix B for a full description of the parameter settings). We ran the simulation two ways, once without criterial variance and once with criterial variance (see Appendix B for the values of criterial variance). Again, we do not believe that the model that assumes no criterial variance (i.e., assumes that every witness had exactly the same decision criteria) is a credible model. We included the no-criterial-variance model here to underscore that the assumption of no variance has important consequences despite the fact that computational modeling of eyewitness identification processes has failed to include criterial variance (e.g., Clark, 2003; Clark, Erickson, & Breneman, 2011; Wixted & Mickes, 2015b).

The result of the simulation that includes criterial variance is shown in Figure 3. It is clear that the fair lineup produces a higher ROC curve than the biased lineup. And this higher ROC curve occurs despite the fact that the fair lineup does not have higher discriminability than the biased lineup (it actually has lower discriminability). Instead, the

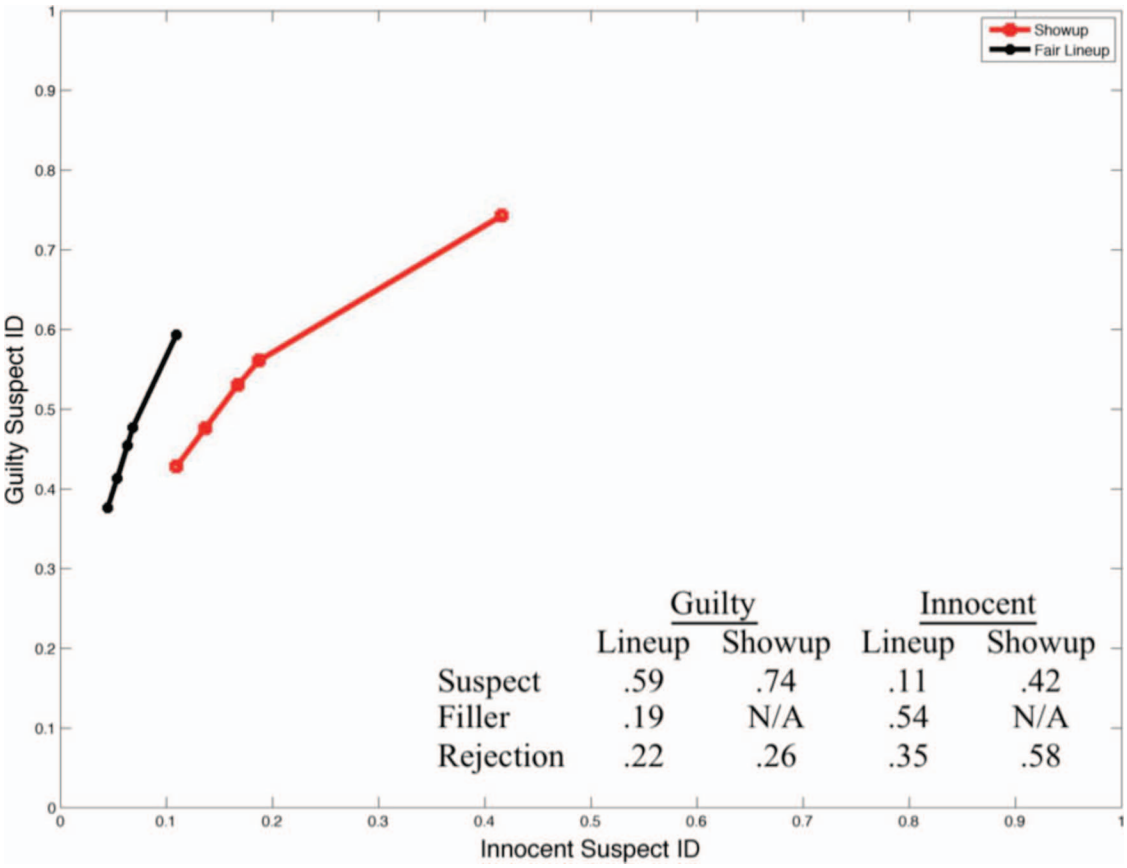


Figure 2. Computational simulation comparing showups (red square markers) to fair lineups (black circular markers) when decision criteria are assumed to vary among eyewitnesses. See the online article for the color version of this figure.

entire effect is the result of differential filler siphoning. Specifically, good fillers siphon more choices away from the innocent suspect than they do from the culprit whereas poor fillers siphon few choices at all, regardless of whether the suspect is innocent or guilty. In fact, the fair lineup produced more false alarms than the biased lineup (when false alarms include both the innocent suspect and all false identifications of fillers) and the fair lineup also produced fewer identifications of the culprit than did the biased lineup. Therefore, as Wells et al. (2015b, 2015c) had previously noted from Wetmore et al.'s (2015) data on fair versus biased lineups, when filler selections are properly treated as false alarms, the fair lineup tends to show lower, not higher, discriminability than does the biased lineup. However, the ROC curve is, nevertheless, higher for the fair lineup because the ROC curve only measures PPV, which is higher for fair lineups, because of differential filler siphoning.

Figure 4 shows what happens when the computational simulation assumes that every witness has exactly the same decision criteria. Figure 4 data do not look at all like what one gets from actual witnesses. It takes a proper model to simulate the differential filler siphoning effect, a model that includes criterial variance. We encourage those who are building computational models of eyewitness identification to include criterial variance. We showed here, for instance, that conclusions about both the lineup versus showup and the fair versus biased lineup changed once we included the assumption that not every witness uses the same exact decision criteria.

More generally, we have shown is that higher ROC curves can be obtained in the absence of any changes in underlying discriminability. In fact, we have shown that the procedure with lower overall discriminability can actually produce a higher ROC curve. This higher ROC for a procedure with inferior discriminability is the synergistic outcome of two phenomena: (a) the use of good fillers decreases discriminability, but also decreases innocent suspect identifications to a greater extent than culprit identifications, resulting in increased PPV, and (b) ROC analysis on lineups is a measure of PPV and not discriminability.

For both of the fundamental problems in eyewitness identification, namely lineups versus showups and fair lineups versus biased lineups, we have shown a clear dissociation between underlying discriminability and the area under an ROC curve. This occurs because ROC analyses on lineups treat false-affirmative identifications of fillers as if they were rejections. The theoretical consequences of this reliance on ROC curves are immense as they lead to unsubstantiated speculation about discriminability mechanisms (such as articulated in Diagnostic-Feature Detection theory) and fail to recognize the simple role of differential filler siphoning.

Why does adding criterial-variance reverse the ordinal ranking of lineup and showup ROC curves in computational simulations? When researchers fit data to computational models in which decision criteria are assumed to be static, filler siphoning occurs, but it is not differential. In other words, when researchers

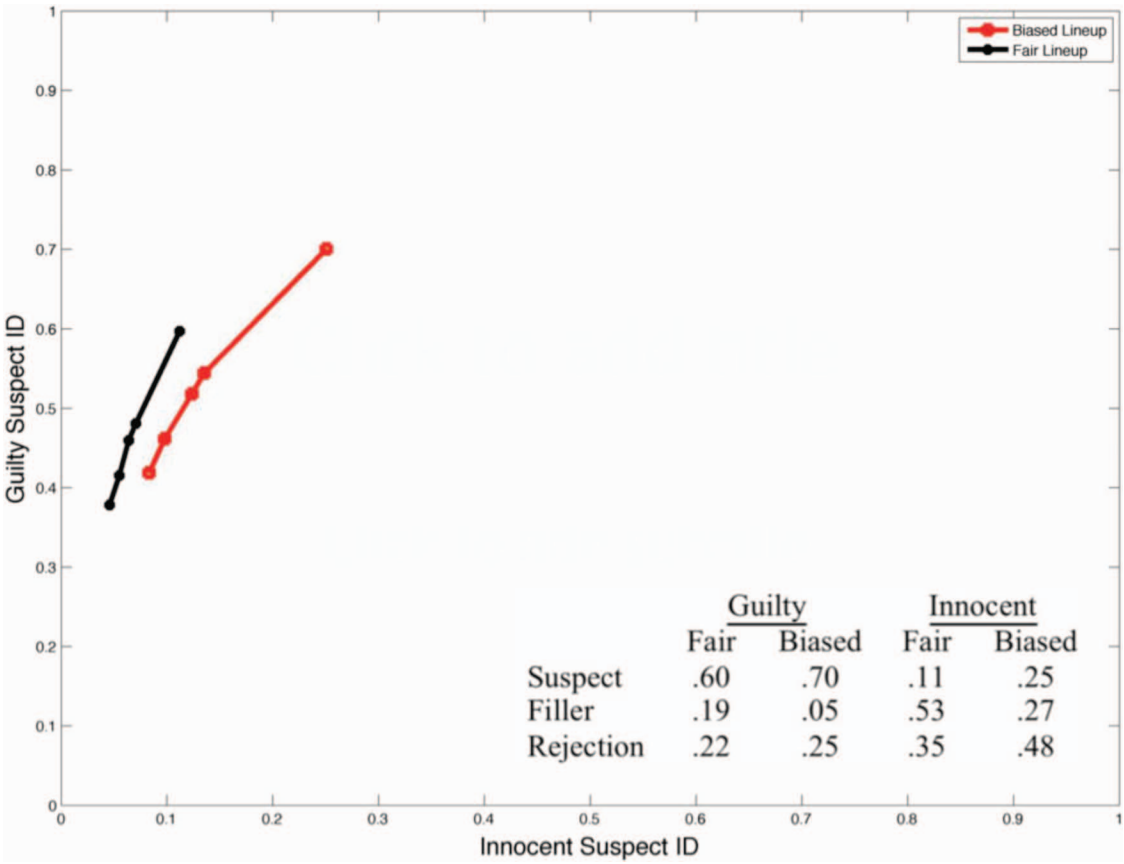


Figure 3. Computational simulation comparing biased lineups (red square markers) to fair lineups (black circular markers) when decision criteria are assumed to vary among eyewitnesses. See the online article for the color version of this figure.

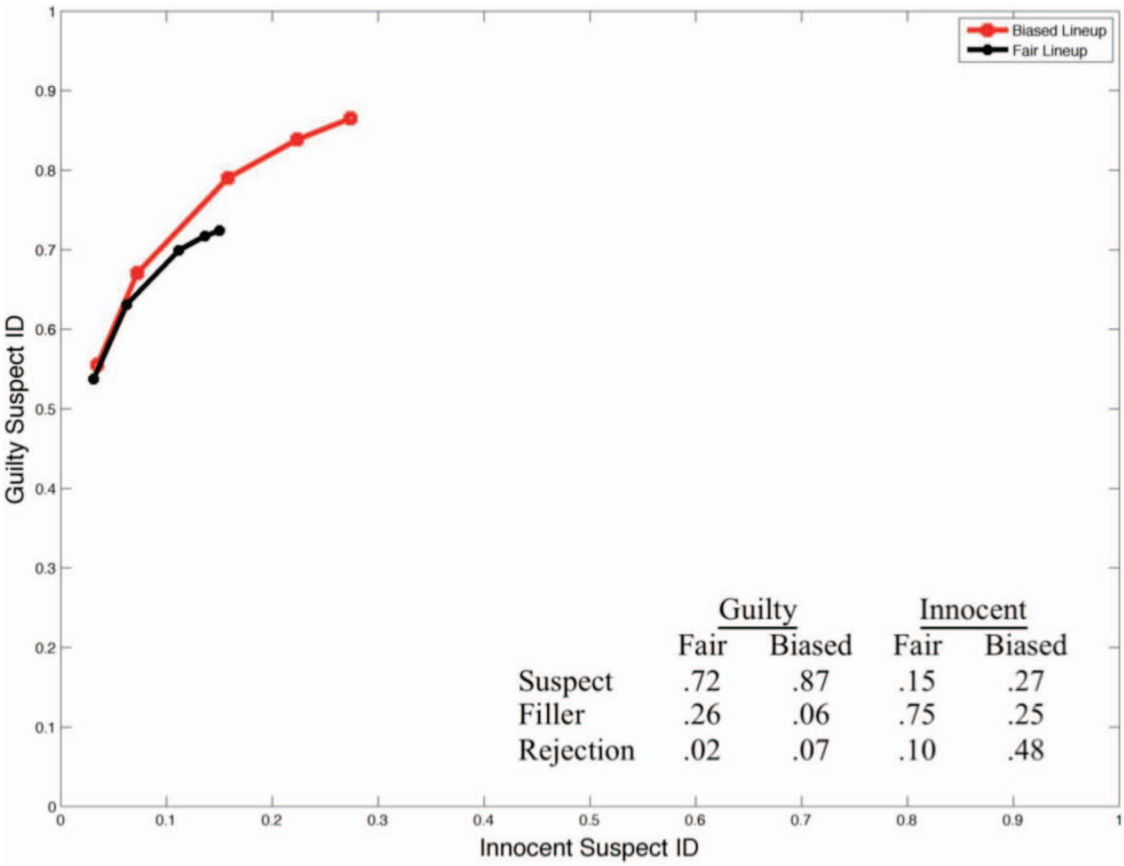
assume that every eyewitness constructs the exact same decision criteria, and that d' on suspect identifications is the same for both fair lineups and showups or fair lineups and biased lineups, fillers siphon as many identifications away from the culprit as they do from the innocent suspect. However, when more realistic models that incorporate criterial variance are used, filler siphoning is differential—fillers siphon more identifications away from the innocent suspect than from the culprit. Why is it that permitting criterial variance results in differential filler siphoning?

When computational models assume that decision criteria are static, they also implicitly assume that the probability that one lineup member will exceed the eyewitness' decision criterion is independent of the probability that any other lineup member will exceed criterion. In other words, when computational models of lineup procedures assume static criteria, they also assume statistical independence. However, the assumption of statistical independence is of course violated in a lineup procedure. Where an eyewitness places his or her decision criterion impacts the probability of identification for all lineup members. For example, if John constructs a decision criterion that is relatively lenient, this increases every lineup members' probability of being identified. Likewise, if Laura constructs a decision criterion that is relatively stringent, this decreases every lineup members' probability of being identified.

Why does violating the assumption of statistical independence matter? Consider again the data in Figures 2 and 3. Under the assumption of static decision criteria (see Figure 2), the showup resulted in a culprit identification rate of 89% and an innocent suspect identification rate of 32%. However, when criterial variance is permitted (see Figure 3), the culprit identification rate decreased to 74% and the innocent suspect identification rate increased to 42%. For the showup procedure, the addition of criterial variance produced a classic "mirror effect" as culprit identifications decreased and innocent suspect identifications increased, *mirroring* the pattern observed under the assumption of static criteria.

Under the assumption of static criteria (see Figure 2), the lineup resulted in a culprit identification rate of 72%, a culprit-present filler identification rate of 26%, and a culprit-absent false-positive rate of 90% (innocent suspect identification rate = 90%/6 lineup members = 15%). As in the showup procedure, permitting criterial variance decreased the culprit identification rate in the lineup procedure (to 59%) and also decreased culprit-present filler identifications to 19%. Most importantly, adding criterial variance *decreased* the culprit-absent false-positive rate to 65% (innocent suspect identification = 10.83%).

Unlike the showup procedure in which adding criterial variance increased both false-positive and false-rejection errors, adding



	Guilty		Innocent	
	Fair	Biased	Fair	Biased
Suspect	.72	.87	.15	.27
Filler	.26	.06	.75	.25
Rejection	.02	.07	.10	.48

Figure 4. Computational simulation comparing biased lineups (red square markers) to fair lineups (black circular markers) when it is assumed that every witness uses exactly the same decision criterion. See the online article for the color version of this figure.

criterial variance to the lineup increased false-rejection errors but decreased false-positive errors. To understand why this happened, consider Panels A and B of Figure 5. In both panels, the thick black normal distribution represents the range of familiarity values for the culprit, the solid gray normal distribution represents the range of familiarity values for the innocent suspect, and the dashed gray normal distribution represents the familiarity values for each of the five fillers. The black vertical line in Panel A is an unbiased decision criterion for the static criterion model and the relatively platykurtic normal distribution in Panel B is an unbiased decision criterion for the criterial variance model.

The first point to address is why adding criterial variance decreases culprit identifications for the lineup and showup procedures. The part of the criterion distribution that falls below the mean represents eyewitnesses with relatively lenient criteria and the part of the distribution that falls above the mean represents eyewitnesses with relatively stringent criteria. On the one hand, only a relatively small proportion of the culprit distribution falls below the mean of the criterion distribution. Thus, there is little opportunity for eyewitnesses with relatively lenient decision criteria to increase the culprit identification rate. On the other hand, a relatively large proportion of the culprit distribution falls above the mean of the criterion distribution. Thus, there is ample opportunity for eyewitnesses with relatively stringent decision criteria to decrease the culprit identification rate. Taking these two patterns together, adding criterial variance decreases culprit identifications for both lineup and showup procedures.

Things get more interesting when considering how criterial variance impacts false-affirmative responses in lineup and showup procedures. Only a relatively small proportion of the innocent suspect (and good filler) distribution exceeds the mean of the criterion distribution. Thus, there is little opportunity for eyewitnesses with relatively stringent decision criteria to increase correct rejections. In contrast, a large proportion of the innocent suspect (and good filler) distribution falls below the mean of the criterion distribution. Thus, there is ample opportunity for eyewitnesses with relatively lenient decision criteria to increase false-positive identifications. For the showup procedure, the result is an increase in innocent suspect identifications. However, adding criterial variance to the lineup procedure actually leads to an increase in correct rejections.

The increase in correct rejections that comes from adding criterial variance to lineup procedures can be explained by considering two factors. First, under the assumption that the familiarity of lineup members is statistically independent (viz., the static criteria model), 90% of simulated-eyewitnesses who encountered a culprit-absent procedure made a false-affirmative identification. Thus, criterial variance can only possibly increase the false-

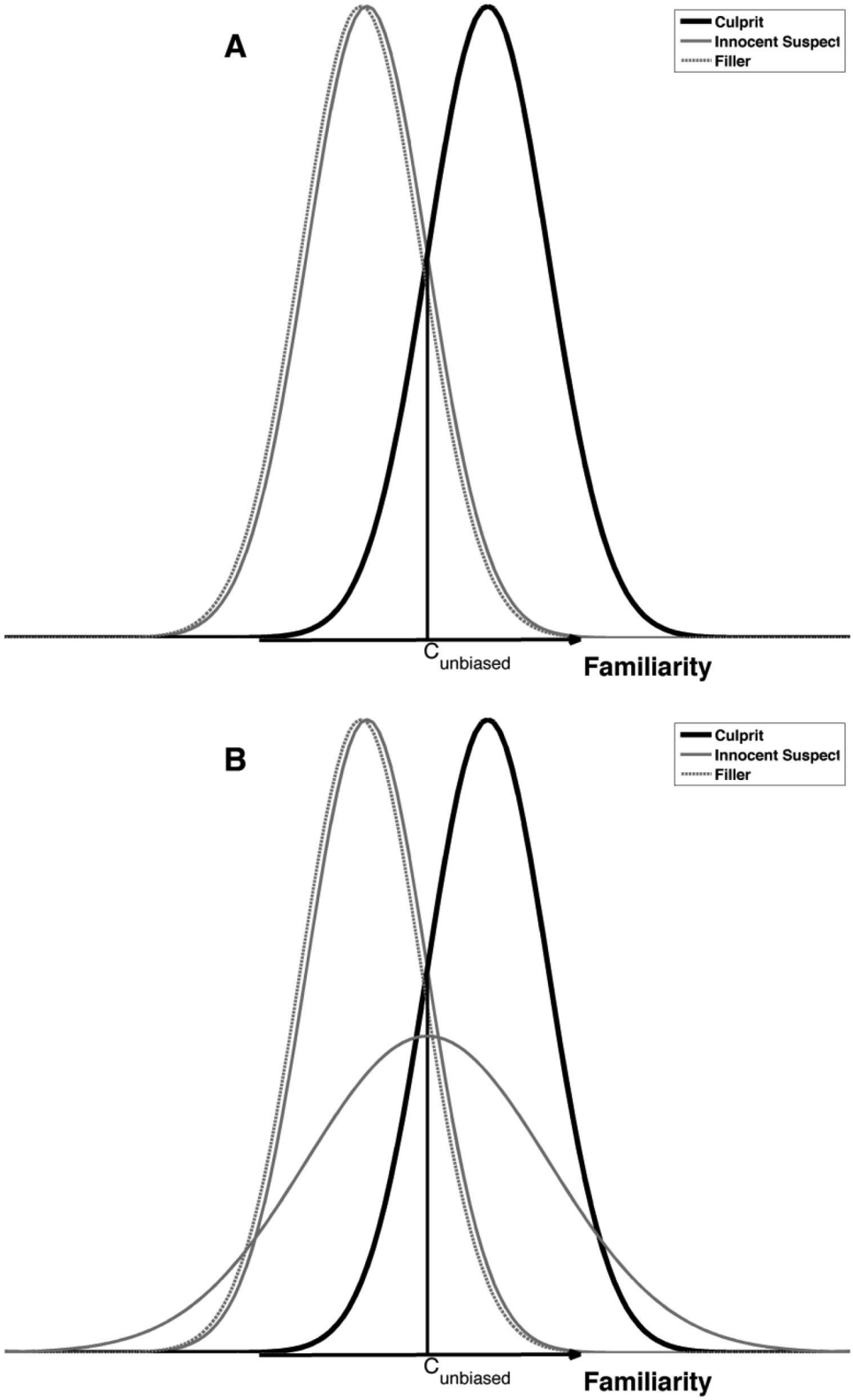


Figure 5. Panel A represents a fair lineup procedure under the assumption of static criteria and Panel B represents a fair lineup procedure under the assumption of criterial variance. In both instances, one can ignore the filler distributions to draw inferences about showup procedures.

affirmative identification rate by a maximum of 10%. Second, although only a relatively small proportion of the innocent suspect distribution exceeds the mean of the criterion distribution, there are actually six such distributions—one for the innocent suspect and one for each of the five fillers. To determine the extent to which criterial variance might increase correct rejections, we must multiply the proportion of the innocent suspect distribution that exceeds the criterion distribution by six. The result is that there is ample opportunity for criterial variance to increase correct rejections but little opportunity to increase false alarms. Thus, when permitting criterial variance in lineup procedures, correct rejections increase.

Signal Detection Theory Naturally Predicts That Fair Lineups Will Increase PPV Despite Decreasing Underlying Discriminability

For this section we focus on the comparison of fair and biased lineup, but given that showups are essentially the extreme of biased lineups, all of the points we make here are equally applicable to the comparison of fair lineups to showups. Consider the two panels in Figure 6. Panel A is a biased lineup and Panel B is a fair lineup. Panel B is identical to the criterial variance model presented in Panel B of Figure 5. Panel A of Figure 6 is also identical to these lineups with the exception that it is biased—the innocent suspect better resembles the culprit than do the fillers.

Examining the simple SDT models in Figure 6, it is readily apparent that SDT naturally predicts fair lineups to result in worse discriminability than biased lineups, but to have superior PPV. First consider the underlying discriminability associated with the two procedures. When the culprit is absent, SDT predicts that the fair lineup will result in more total false-affirmative responses (innocent suspect plus fillers) than the biased lineup. In both instances, the same proportion of the innocent suspect distribution exceeds the decision criterion, but in the fair lineup, a much larger proportion of the filler distribution exceeds criterion. Thus, SDT predicts that fair lineups will produce more false alarms than biased lineups and computational simulations (see Figure 3) and empirical data (Wetmore et al., 2015) bear out that prediction. When the culprit is present, SDT predicts that the fair lineup will result in fewer culprit identifications than the biased lineup. This is evident from the fact that the good filler distribution (Panel B) overlaps the culprit distribution to a greater extent than the poor filler distribution (Panel A). Thus, good fillers will siphon more identifications away from the culprit than will poor fillers, and both computational simulations (see Figure 3) and empirical data (Wetmore et al., 2015) bear out that prediction as well. This also makes intuitive sense. By increasing the extent to which fillers resemble a suspect (or by increasing the number of fillers in a lineup), one is adding noise to the identification procedure and as noise increases, it becomes increasingly difficult to discriminate between previously seen and novel faces. The result of increasing noise of course is more identification errors.

There is no way to reconcile the finding that biased lineups result in more culprit identifications and fewer false-affirmative identifications than fair lineups with the idea that fair lineups have superior underlying discriminability. It is objectively clear in both computational simulations and empirical data that biased lineups have better, not worse, discriminability than fair lineups. Furthermore, as we have shown in Figure 6, this is precisely what SDT predicts and as we demonstrated above, it is precisely what a fit of the SDT-CD model to the empirical data shows.

It is also readily apparent from Figure 6 that SDT naturally predicts fair lineups to have superior PPV when compared with biased lineups because of differential filler siphoning. SDT predicts that good fillers will siphon more identifications away from both the culprit and the innocent suspect than will poor fillers; but, SDT also predicts that the increase in siphoned identifications will be greater when the suspect is innocent than when the suspect is guilty. The fact that good fillers will siphon more identifications away from suspects in general is evident from the fact that the good filler distribution (Panel B) overlaps both the culprit and innocent suspect distributions to a greater extent than the poor filler distribution (Panel A). Because switching from poor to good fillers increases the overlap between the filler and innocent suspect distributions to a greater extent than it increases the overlap between filler and culprit distributions, it is evident that switching to good fillers will decrease innocent suspect identifications to a greater extent than culprit identifications. And, as was the case with our discussion of underlying discriminability, these theoretical predictions are also reflected in computational simulations (see Figure 3) and empirical data (Wetmore et al., 2015).

Conclusions

We addressed two primary assumptions underlying the recommendation that ROC analysis be used to compare lineups: (a) ROC analysis on lineups measures underlying discriminability, and (b) the procedure with superior underlying discriminability is the procedure that produces superior applied utility. These assumptions were also used to derive a psychological processing theory intended to explain why lineups are superior to showups and why fair lineups are superior to biased lineups (Diagnostic-Feature Detection; Wixted, & Mickes, 2014). In the present article, we have demonstrated that ROC analysis on lineups does not measure underlying discriminability and that, despite intuition, the procedure that promotes superior underlying discriminability is not necessarily the procedure that produces superior forensic outcomes. We have also raised serious concerns about referring to ROC analysis on lineups as a measure of applied utility or objective discriminability. Because ROC analysis on lineups reflects only PPV and not NPV or any form of discriminability, it should not be referred to as a measure of applied utility or objective discriminability. Rather, ROC analysis on lineups should be referred to for what it is, a measure of PPV. Because ROC analysis on lineups treats false-affirmative filler identifications as correct rejections it cannot assess the underlying discriminability associated with lineup procedures. The assumption that ROC analysis on lineups measures underlying discriminability led Wixted and Mickes (2014) to posit a theory about how surrounding suspects with good fillers improves discriminability. As we have demonstrated in this article, discriminability is not greater for lineups than for showups and discriminability is not better for fair lineups than for biased lineups. Instead, a completely structural phenomenon, differential filler siphoning, explains why lineups are superior to showups and why fair lineups are superior to biased lineups. Differential filler siphoning does not arise from an increase in underlying discriminability, but instead arises from simply spread-

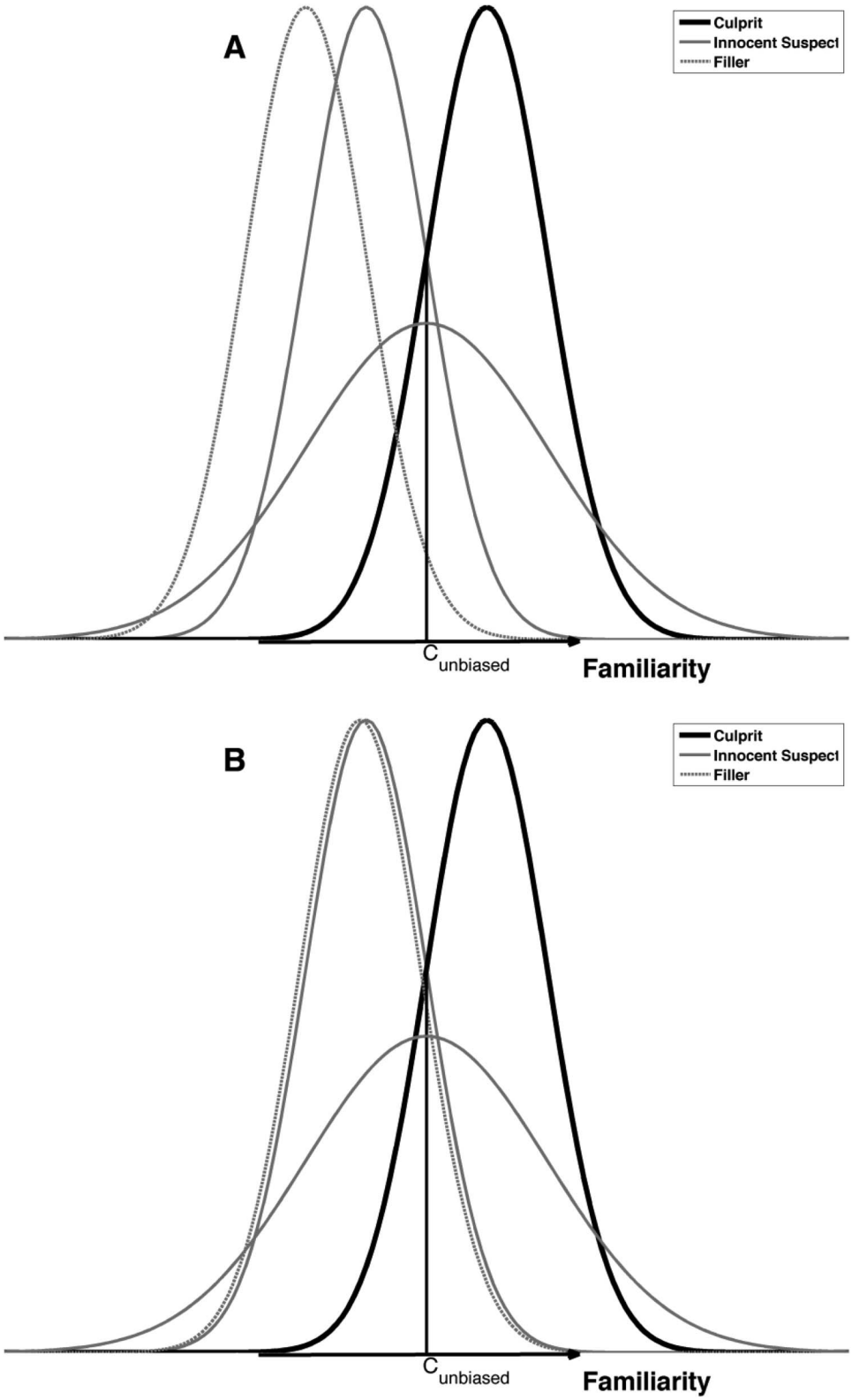


Figure 6. Panel A represents a biased lineup procedure and Panel B represents a fair lineup procedure.

ing affirmative mistaken identifications away from the innocent suspect and toward fillers.

We also pointed out that Wixted and Mickes (2015b) reached the wrong conclusion when they argued that their computational models required a memorial advantage for lineups over showups to produce a higher ROC for lineups. In fact, their computational simulation required higher PPV for lineups than for showups, but not a memorial advantage. Wixted and Mickes (2015b) reached the wrong conclusion on this point because they assumed a d' score that focused solely on suspect identifications was a measure of underlying discriminability. We have demonstrated that PPVd' is not a measure of underlying discriminability, but is a measure of PPV.

In addition, we demonstrated that the untenable assumption that all eyewitnesses share the same decision criteria are yet another reason past computational simulations have failed to find an ROC advantage for fair lineups relative to showups (e.g., Wixted & Mickes, 2015b). Once criterial variance is permitted, simulations comparing fair lineups and showups display an ROC advantage for fair lineups even when d' on suspect identifications is held constant (see Figure 2). This is precisely the pattern of results that is predicted by realistic SDT models of lineup procedures (see Figure 6). In other words, by not making the assumption that decision criteria are nonvariable among eyewitnesses, we corrected the pattern of results that Wixted and Mickes (2015b) reported comparing fair lineups and showups. We then showed that this same pattern emerges for comparisons of fair lineups to biased lineups. This finding is of importance for demonstrating that differential filler siphoning and not Diagnostic-Feature Detection can explain why fair lineups produce a higher ROC curve than do showups, but there are at least two additional widespread implications of this finding. First, this finding adds to growing evidence documenting the important role criterial variability can play in explaining experiments and simulations (e.g., Benjamin et al., 2009; Mueller & Weidemann, 2008). Second, this finding suggests that some previous conclusions about eyewitness identification based on computational modeling (e.g., the WITNESS model; Clark, 2003) might need to be re-examined with models that include criterial variance.

It might be surprising to many eyewitness researchers to learn that using good lineup fillers (a fair lineup) actually reduces discriminability relative to poor fillers (a biased lineup) or no fillers (a showup). However, in fact one of the one of the most fundamental findings in recognition memory research is that increasing the similarity of lures to the target reduces memory discriminability. It was also probably surprising to many that an identification procedure can produce worse underlying discriminability relative to some other identification procedure, yet have superior PPV. But, as we have demonstrated through analyzing recent data (e.g., Wetmore et al., 2015), and as is evidenced by our simple SDT model of differential filler siphoning, other processes (i.e., differential filler siphoning) can have a greater impact on PPV than does discriminability. To be clear, all else being equal, the procedure with superior discriminability will have superior PPV. However, processes other than discriminability are at work in lineup procedures and one cannot assume that the reason one procedure has superior PPV relative to some other procedure is because it better enhances discriminability. We conclude with a few general points. First, we would argue that ROC analysis on lineups should no longer be referred to as a measure of any kind of discriminability. ROC analysis should be accurately referred to for what it is: a measure of PPV.

Second, we acknowledge that there is no single way to analyze data from lineup identification procedures (cf. Gronlund, Wixted, & Mickes, 2014). However, regardless of the method used in any given experiment, it should be incumbent on researchers to report data from the full 3×2 array instead of focusing myopically on just the suspect identification data. We have shown in the current article the critical role played by fillers in understanding a core process that explains the results. Finally, although we think that Bayesian measures of diagnosticity have a lot to offer and avoid many of the pitfalls associated with ROC analysis on lineups, Bayesian measures have limitations. The diagnosticity of suspect identifications (culprit identification/innocent suspect identification) does increase monotonically with increasingly conservative responding (Wixted & Mickes, 2014). However, if one examines all six cells from a lineup, it is evident that the NPV associated with a lineup procedure monotonically decreases with increasingly conservative responding (Wells et al., 2015a). In other words, if a procedure increases the diagnosticity of suspect identifications simply because it exacts more conservative responding, this procedure will pay a cost in that the diagnosticity of exonerating behaviors will decrease. However lineup researchers analyze their data, they must consider suspect identifications, filler identifications, and rejections. Failure to consider all three eyewitness behaviors will stymy both theoretical and applied developments.

References

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115. http://dx.doi.org/10.1037/a0014351

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17,* 629–654. http://dx.doi.org/10.1002/acp.891

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238–259. http://dx.doi.org/10.1177/1745691612439584

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35,* 364–380. http://dx.doi.org/10.1007/s10979-010-9245-1

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16,* 22–42. http://dx.doi.org/10.3758/PBR.16.1.22

Cutler, B. L. (2013). *Reform of eyewitness identification procedures*. Washington, DC: American Psychological Association Press. http://dx.doi.org/10.1037/14094-000

Duncan, M. J. (2006). *A signal detection model of compound decision tasks* (Tech. Rep. Np. No. TR2006-256). Toronto, ON, Canada: Defence Research and Development Canada.

Egan, J. P. (1958). *Recognition memory and the operating characteristic*. (Technical Note AFCRC-TN_58-51, AO-152650). Bloomington, IN: Indiana University Hearing and Communication Laboratory.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Wiley.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science, 23,* 3–10. http://dx.doi.org/10.1177/0963721413498891

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in Signal Detection Theory: The case of recognition memory. *Psychological Review, 119,* 457–479. http://dx.doi.org/10.1037/a0027727

Kellen, D., Klauer, K. C., & Singmann, H. (2013). On the measurement of criterion noise in Signal Detection Theory: Reply to Benjamin (2013). *Psychological Review, 120,* 727–730. http://dx.doi.org/10.1037/a0033141

Lampinen, J. M. (2016). ROC analysis in eyewitness identification research. *Journal of Applied Research in Memory & Cognition, 5,* 21–33. http://dx.doi.org/10.1016/j.jarmac.2015.08.006

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376. http://dx.doi.org/10.1037/a0030609

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute *d',* not the diagnosticity ratio. *Journal of Applied Research in Memory & Cognition, 3,* 58–62. http://dx.doi.org/10.1016/j.jarmac.2014.04.007

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865. http://dx.doi.org/10.3758/BF03194112

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494. http://dx.doi.org/10.3758/PBR.15.3.465

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

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. http://dx.doi.org/10.1037/h0093923

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied, 16,* 387–398. http://dx.doi.org/10.1037/a0021034

Pryke, S., Lindsay, R. C. L., Dysart, J. E., & Dupuis, P. (2004). Multiple independent identification decisions: A method of calibrating eyewitness identifications. *Journal of Applied Psychology, 89,* 73–84. http://dx.doi.org/10.1037/0021-9010.89.1.73

Starr, S. J., Metz, C. E., Lusted, L. B., & Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiology, 116,* 533–538. http://dx.doi.org/10.1148/116.3.533

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 27,* 523–540. http://dx.doi.org/10.1023/A:1025438223608

Wells, G. L. (2001). Police lineups: Data, theory, and policy. *Psychology, Public Policy, and Law, 7,* 791–801. http://dx.doi.org/10.1037/1076-8971.7.4.791

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88,* 776–784. http://dx.doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory & Cognition, 4,* 313–317. http://dx.doi.org/10.1016/j.jarmac.2015.08.008

Wells, G. L., Smith, A. M., & Smalarz, L. (2015). ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory & Cognition, 4,* 324–328. http://dx.doi.org/10.1016/j.jarmac.2015.08.010

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior, 39,* 99–122. http://dx.doi.org/10.1037/lhb0000125

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory & Cognition, 4,* 8–14. http://dx.doi.org/10.1016/j.jarmac.2014.07.003

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7,* 275–278. http://dx.doi.org/10.1177/1745691612442906

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262–276. http://dx.doi.org/10.1037/a0035940

Wixted, J. T., & Mickes, L. (2015a). Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory & Cognition, 4,* 318–323. http://dx.doi.org/10.1016/j.jarmac.2015.08.009

Wixted, J. T., & Mickes, L. (2015b). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory & Cognition, 4,* 329–334. http://dx.doi.org/10.1016/j.jarmac.2015.08.007

Appendix A

Computational Simulations Comparing Fair Lineups and Showups

The computational simulations for the ROC curves shown in Figures 2 and 3 were generated with a simulation based on the "best-match" strategy of decision-making in lineup procedures (Clark, 2003; Lampinen, 2016). The best-match strategy has two steps. First, the eyewitness compares all lineup members and determines which lineup member is the most familiar. Second, the eyewitness compares the most familiar lineup member to his or her decision criterion and if the lineup member exceeds that criterion, the eyewitness makes an identification. We chose the best-match strategy because it is the de facto strategy in showup procedures (i.e., the eyewitness simply compares the lone suspect to his or her decision criteria) and if we want to attribute any difference in the showup and lineup ROCs to the presence of fillers, then it is necessary that we hold all other parameters constant.

The fair lineups in our simulation included one suspect and five fillers. The showups in our simulation included only a single suspect and no fillers. We assumed that all members in our identification procedures evoke some level of familiarity and when present, the culprit evokes a higher degree of familiarity, on average, than fillers. The mean difference between these distributions is d'. We assumed that all fillers in a given condition evoked the same familiarity, on average. In culprit-absent lineups, we assumed that all lineup members were drawn from the same distribution (i.e., the innocent suspect is no more familiar than the fillers). We had five such decision criteria: C1 = 0.48, C2 = 0.69, C3 = 0.97, C4 = 1.45, and C5 = 1.83. If the familiarity of all lineup members was below C1, the lineup was rejected. If the familiarity of at least one lineup member exceeded C1, the most familiar lineup member was identified. Confidence was equal to the largest decision criterion that the familiarity value exceeded.

We used the unequal variance SDT model proposed by Wixted and Mickes (2014; see also Mickes, Wixted, & Wais, 2007). For each culprit-present lineup we randomly drew one value from the culprit distribution ($\mu = d'$, $\sigma = 1.22$) and five values from the filler distribution ($\mu = 0$, $\sigma = 1.00$). For each culprit-absent lineup, we randomly drew six values from the filler distribution ($\mu = 0$, $\sigma = 1.00$), one of which was designated as our innocent suspect. For each culprit-present showup, we randomly drew one value from the culprit distribution ($\mu = d'$, $\sigma = 1.22$). For each culprit-absent showup, we randomly drew one value from the innocent suspect distribution ($\mu = 0$, $\sigma = 1.00$).

Each of the four curves depicted in Figures 2 and 3 was based on 50,000 observations. For the simulation depicted in Figure 2, we assumed (as in Wixted and Mickes, 2015b) that every witness had exactly the same decision criterion (i.e., the absence of criterial variance). For the simulation depicted in Figure 3, however, we assumed that decision criteria vary among eyewitnesses (viz. there is criterial variance). Hence, in Figure 3, each decision criterion was drawn from a normal distribution with $\mu = C_i$, $\sigma = 2.00$. Note that, because these decision criteria are probabilistic, they do not necessarily maintain ordinal ranking within-eyewitness. For example, on any given trial, the C4 criterion could be more stringent than the C5 criterion. In appreciation of this, we ran additional simulations in which we constrained decision criteria to maintain ordinal ranking and we obtained precisely the same results as when we did not use such a constraint.

Appendix B

Computational Simulations Comparing Biased Lineups and Showups

The computational simulations for the ROC curves shown in Figures 4 and 5 were generated the same way as those in Figures 2 and 3 (see Appendix A) with regard to the best match model, unequal variance SDT, and so on. The fair lineup was generated with the same parameter settings as those used in Figures 2 and 3, as described in Appendix A. The biased lineup used all these same parameters except that five values from the filler distribution were set to ($\mu = -1$, $\sigma = 1.00$).

Received April 28, 2016
Revision received August 6, 2016
Accepted August 23, 2016 ■