Distinguishing Between Investigator Discriminability and Eyewitness Discriminability: A Method for Creating Full Receiver Operating Characteristic Curves of Lineup Identification Performance

Andrew M. Smith
*Iowa State University, amsmith@iastate.edu*

Yueran Yang
*University of Nevada, Reno*

Gary L. Wells
*Iowa State University, glwells@iastate.edu*

Follow this and additional works at: [https://lib.dr.iastate.edu/psychology_pubs](https://lib.dr.iastate.edu/psychology_pubs)

Part of the Cognitive and Perception Commons, Experimental Analysis of Behavior Commons, Human Factors Psychology Commons, Law and Psychology Commons, Public Law and Legal Theory Commons, and the Theory and Philosophy Commons

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/psychology_pubs/96](https://lib.dr.iastate.edu/psychology_pubs/96). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Distinguishing Between Investigator Discriminability and Eyewitness Discriminability: A Method for Creating Full Receiver Operating Characteristic Curves of Lineup Identification Performance

Abstract
The conceptual frameworks provided by both the lineups-as-experiments analogy and Signal Detection Theory have proven important to furthering understanding of performance on eyewitness identification procedures. The lineups-as-experiments analogy proposes that when investigators carry out a lineup procedure, they are acting as experimenters, and should therefore follow the same tried-and-true procedures that experimenters follow when executing an experiment. Signal Detection Theory offers a framework for distinguishing between factors that improve the trade-off between culprit and innocent-suspect identifications (discriminability) and factors that impact the frequency of suspect identifications (conservativeness). The present work offers an integration of these two conceptual frameworks. We argue that an eyewitness lineup procedure is characterized by two simultaneous Signal Detection tasks. On one hand, the witness is tasked with determining whether the culprit is present in the lineup and whom that person is. On the other hand, the investigator knows which lineup member is the suspect and which lineup members are known-innocent fillers and is therefore tasked only with determining whether the suspect is the culprit. The investigator uses the witness' identification decision and associated level of confidence to make a decision about whether the suspect is the culprit. We leverage this realization to demonstrate a method for creating full Receiver Operating Characteristic (ROC) curves for eyewitness lineup procedures and demonstrate that the conclusions drawn from comparing full lineup ROC curves differ from those drawn from comparing suspect-only partial ROC curves.

Keywords
Eyewitness Lineups, ROC Analysis, Signal Detection Theory, Investigator Discriminability

Disciplines
Cognition and Perception | Experimental Analysis of Behavior | Human Factors Psychology | Law and Psychology | Public Law and Legal Theory | Theory and Philosophy

Comments

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/psychology_pubs/96
Distinguishing Between Investigator Discriminability and Eyewitness Discriminability: A Method for Creating Full Receiver Operating Characteristic Curves of Lineup Identification P...
Distinguishing Between Investigator Discriminability and Eyewitness Discriminability: A Method for Creating Full Receiver Operating Characteristic Curves of Lineup Identification Performance

Andrew M. Smith, Yueran Yang, & Gary L. Wells

Author Note
Andrew M. Smith, Department of Psychology, Iowa State University; Yueran Yang, Department of Psychology, University of Nevada Reno; Gary L. Wells, Department of Psychology, Iowa State University.
This work was supported by a grant from Arnold Ventures, 400-61-12C. The views and opinions expressed in this work are those of the authors and do not necessarily reflect the views of Arnold Ventures.
Correspondence concerning this article should be addressed to Andrew M. Smith, Department of Psychology, W112 Lagomarcino Hall, Iowa State University, 50011.
E-mail: amsmith@iastate.edu
Abstract

The conceptual frameworks provided by both the lineups-as-experiments analogy and Signal Detection Theory have proven important to furthering understanding of performance on eyewitness identification-procedures. The lineups-as-experiments analogy proposes that when investigators carry out a lineup procedure, they are acting as experimenters, and should therefore follow the same tried-and-true procedures that experimenters follow when executing an experiment. Signal Detection Theory offers a framework for distinguishing between factors that improve the trade-off between culprit and innocent-suspect identifications (discriminability) and factors that impact the frequency of suspect identifications (conservativeness). The present work offers an integration of these two conceptual frameworks. We argue that an eyewitness lineup procedure is characterized by two simultaneous Signal Detection tasks. On one hand, the witness is tasked with determining whether the culprit is present in the lineup and whom that person is. On the other hand, the investigator knows which lineup member is the suspect and which lineup members are known-innocent fillers and is therefore tasked only with determining whether the suspect is the culprit. The investigator uses the witness' identification decision and associated level of confidence to make a decision about whether the suspect is the culprit. We leverage this realization to demonstrate a method for creating full Receiver Operating Characteristic (ROC) curves for eyewitness lineup procedures and demonstrate that the conclusions drawn from comparing full lineup ROC curves differ from those drawn from comparing suspect-only partial ROC curves.

*Keywords:* Eyewitness Lineups; ROC Analysis; Signal Detection Theory; Investigator Discriminability
INVESTIGATOR PERFORMANCE VERSUS WITNESS PERFORMANCE

Distinguishing Between Investigator Discriminability and Eyewitness Discriminability: A Method for Creating Full Receiver Operating Characteristic Curves of Lineup Identification Performance

Over the past 40 years, psychological scientists have learned a great deal about the factors that impact performance on eyewitness identification procedures. Two theoretical foundations have proven particularly useful: Wells and Luus' (1990) lineups-as-experiments analogy and Signal Detection Theory (e.g., Clark, 2003; Palmer & Brewer, 2012). The lineups-as-experiments analogy proposes that when police investigators carry out a lineup procedure, they are acting as experimenters, testing the hypothesis that their suspect is the culprit. By extension, investigators should craft their lineups with the same care and based on the same principles that scientists employ when crafting an experiment. This analogy has led to many fruitful developments in how investigators construct and administer lineup procedures. At the other end of the spectrum, Signal Detection Theory has largely been used to examine how differences in lineup procedures and variations in memory quality affect the trade-off between culprit and innocent-suspect identifications. In our view, neither of these theoretical frameworks has yet to come full circle. We argue that during an eyewitness identification procedure, two interrelated, yet distinct, Signal Detection tasks are operating simultaneously. On one hand, the witness is tasked both with determining whether the culprit is present in the lineup and if so, whom that person is. On the other hand, investigators are tasked only with determining whether or not the suspect is the culprit and use the witness' decision, and confidence in that decision, as evidence to make this determination. This distinction proves important both for our understanding of how identification procedures work and
for how we analyze and interpret data from eyewitness experiments. We examine the use of Receiver Operating Characteristic (ROC) curves in the analysis of eyewitness lineups and show that a fundamental misconception about the objective of a lineup procedure led to the precarious development of partial ROC curves. Once we appreciate that it is ultimately the investigator's decision to arrest or release the suspect (based on the information provided by the eyewitness), a method for creating full eyewitness ROC curves becomes readily apparent.

This work was inspired, in part, by a recent call from the National Academy of Sciences (2014) for greater exploration of measures that might be used to assess the quality of eyewitness lineup procedures. At the time of their review, two measures of performance had dominated the eyewitness literature: the diagnosticity ratio (Wells & Lindsay, 1980) and the partial area under the ROC curve (pAUC; Mickes, Flowe & Wixted, 2012; Wixted & Mickes, 2012). The National Academy of Sciences (2014) noted that ROC analysis was an improvement over the diagnosticity ratio if for no other reason than that an ROC curve conveys more information than does the diagnosticity ratio. But, the National Academy of Sciences (2014) also raised several concerns over using the pAUC measure. Since that time, we have demonstrated that there was good reason for trepidation given that the pAUC measure does not provide dispositive information about which of two lineup procedures is superior (Lampinen, Smith, Wells, 2019; Smith, Lampinen, Wells, Smalarz, & Mackovichova, 2019).

In addition to raising concerns about the pAUC measure, the National Academy of Sciences (2014) also criticized eyewitness science for myopically focusing on suspect identifications and largely ignoring other eyewitness behaviors (i.e., filler identifications
and rejections). Examining full ROC curves and the associated full AUC measure offers potential resolution to both of these concerns. Indeed, full ROC curves plot not only suspect identifications but also filler identifications and rejections. Hence, all outcomes from an eyewitness lineup are taken into consideration when deciding which lineup procedure is superior.

The Lineups-As-Experiments Analogy

A lineup is an eyewitness identification procedure in which law-enforcement personnel surround a single suspect with some number (usually five) of known-innocent persons called fillers. The logic for surrounding the suspect with known-innocent fillers is that if the suspect is innocent, fillers offer this individual some protection from mistaken identification in that witnesses will often identify a filler rather than the innocent suspect. Importantly, because fillers are known-innocent persons, these individuals are not at risk of arrest and wrongful conviction. Rather, the inclusion of fillers in lineups creates the possibility for a known error. Because ground truth is unknown in the real world, the opportunity for a known-error is important as it provides law enforcement personnel with some means of distinguishing between reliable and unreliable eyewitnesses. More generally, each of the three possible outcomes from a lineup identification procedure (suspect identification, filler pick, or rejection) holds diagnostic value (Wells & Lindsay, 1980; Wells & Turtle, 1986; Wells, Yang, & Smalarz, 2015). Because suspect identifications occur more frequently when the suspect is guilty than when the suspect is innocent, a suspect identification is diagnostic of guilt. Likewise, because rejections and filler picks occur more frequently when the suspect is innocent than when the suspect is guilty, both of these behaviors are diagnostic of
innocence. Moreover, how diagnostic these behaviors are depends on the confidence an eyewitness places in these decisions. High-confidence suspect identifications are more diagnostic of guilt than are low-confidence suspect identifications (Brewer & Wells, 2006; Wixted & Wells, 2017). Likewise, high-confidence filler picks or rejections are more diagnostic of innocence than are low-confidence filler picks or rejections (Wells et al., 2015).

But in the same sense that the diagnostic value of an experiment depends on the quality of that experiment, so too does the diagnostic value of a lineup outcome. According to the lineups-as-experiments analogy, the investigator conducting the lineup is the experimenter, the witnesses are the participants, the instructions are the experimenter's protocol, the suspect is the stimulus, and who is present in the lineup and where these individuals are positioned is part of the design (Wells & Luus, 1990). Most importantly, police also have a hypothesis, namely that the suspect in the lineup procedure is the person who committed the crime in question. Critically, the goal of a lineup procedure is NOT to test the memory of the eyewitness, but rather to test the hypothesis that the suspect is guilty. The witness provides the datum that the investigators use to draw inferences about the guilt of the suspect. Moreover, just as the inferences that an experimenter draws from an experiment are affected by the validity of that experiment, so too are the inferences that an officer draws from a lineup procedure. With this framework in mind, psychological scientists recommend that investigators include only one suspect per lineup procedure (Wells & Turtle, 1986), prevent the suspect from standing out in the lineup (Clark, 2012; Lindsay & Wells, 1980), caution that the offender might not be present in the lineup (Steblay, 1997, 2013), use double-blind testing (Eisen
et al., 2018; Greathouse & Kovera, 2009; Kovera & Evelo, 2017), and collect a confidence statement at the time of identification (Steblay, Wells, & Douglas, 2014; Wells & Bradfield, 1998). While several other recommendations have been made, there is consensus among eyewitness experts that these "pristine" testing conditions maximize the diagnostic value of eyewitness identification procedures (see Wells et al., 2019, for the current state of affairs).

**Signal Detection Tasks**

Although the conceptual framework provided by the lineups-as-experiments analogy has led to many fruitful developments, it has not yet come full circle. Years after the lineups-as-experiments analogy was introduced to the eyewitness literature, Clark (2003) introduced the Signal Detection framework to the eyewitness literature. Although there is some appreciation of how the memory task completed by the witness is distinct from the performance of a lineup procedure (e.g., Smith, Wells, Lindsay, & Penrod, 2017; Smith, Wells, Smalarz, & Lampinen, 2018), this distinction has not yet been formalized. We believe this is a major oversight. Unequivocally, we should be interested in both the performance of witnesses and investigators (and their lineup procedures)\(^1\); but for applied purposes, we are only interested in the performance of the eyewitness to the extent that it facilitates investigator discriminability. Indeed, regardless of how the witness responds to a lineup procedure, the decision to arrest or release falls in the hands of the investigator.

Consider the example of a radiologist attempting to detect the presence or absence of a malignant tumour in an X-ray. The radiologist is in a similar role to that of a police

\(^1\) Some researchers will undoubtedly argue further that we should be interested in the discriminability of judges, jurors, and other players of the court as they are the ones who formally adjudicate guilt and
investigator and the X-ray is in a similar role to that of the eyewitness. The radiologist examines the X-ray and ultimately makes a decision about the presence or absence of a malignant tumour. The clarity of the X-ray affects the performance of the radiologist. For example, radiologists are better able to discriminate between the presence and absence of malignant tumours when viewing digital X-rays compared to when viewing film-based X-rays (Pisano et al., 2005). To be sure, this is due to the fact that digital X-rays increase the amount of information available to the radiologist. Likewise, the police investigator examines the eyewitness' behaviour (i.e., identification decision and confidence) and ultimately makes a decision about the guilt of the suspect. And, just as the quality of the X-ray affects the performance of the radiologist, the quality of the witness' memory affects the performance of the police investigator. To the extent that the witness' memory is strong, the investigator's task is relatively easy and to the extent that the witness' memory is weak, the investigator's task is relatively difficult. Investigator performance might be further moderated by the procedures that she uses to conduct the lineup procedure. Indeed, these procedures influence witness memory and decision-making, which in turn, impacts the ability of the investigator to distinguish between guilty and innocent suspects. So, the eyewitness is the X-ray machine and its software. The behaviors of the eyewitness in response to the procedure constitute a record that can be analyzed just as the X-ray machine creates an image-format record that can be analyzed.

We suspect that the failure to consider the importance of the investigator's detection task (or to even realize that investigators complete their own detection task) is attributable to the fact that the witness completes a detection task of her own. Alternatively, it is possible that researchers have assumed that the detection task
completed by the witness and the detection task completed by the investigator are one and the same. In the remainder of this section, we describe both the detection task completed by the witness and the detection task completed by the investigator. We demonstrate that, while interrelated, these two detection tasks are distinguishable.

*The Witness' Signal Detection Task*

The detection task completed by the witness conforms to a 3 (witness behaviour: suspect pick, filler pick, rejection) x 2 (culprit: present, absent) confusion matrix. Importantly, investigators do not tell the witness which lineup members are known-innocent fillers and which lineup member is the suspect. Instead, the witness is tasked with determining both whether the culprit is present in the lineup (a detection task), and if so, whom that person is (an identification task). Although the intention of a lineup procedure is not to test the memory of the witness, the witness does complete a memory test. But, the investigators do not know the answer to the test; they do not know if the suspect is guilty or innocent (the entire purpose of the lineup is to gather evidence on the likely guilt of the suspect). So, rather than scoring the performance of the witness on the memory test, the investigators use the witness' memory (and identification decision) to make a decision about whether to arrest or release the suspect.

A Signal Detection model of the witness' detection task is illustrated in Figure 1. The model represents the match between the witness' memory for the culprit and each lineup member with a series of Gaussian distributions. There are three distinct distributions, the guilty-suspect distribution, the innocent-suspect distribution, and the filler distribution. To the extent that the eyewitness has a strong memory for the guilty suspect (i.e., culprit), the guilty-suspect distribution will be shifted to the right of the
innocent-suspect and filler distributions. This makes sense because the witness has actually seen the guilty suspect before and therefore the guilty suspect should tend to provide a better match-to-memory than either an innocent suspect or a known-innocent filler. With a properly constructed lineup, an innocent-suspect should not stand out from the known-innocent fillers. In other words, on average, the innocent suspect should not provide a better match to memory than any of the known-innocent fillers. We make that assumption in this model, but have slightly staggered the innocent-suspect and filler distributions so that the reader can distinguish between them.

Figure 1. The guilty-suspect distribution represents the range of potential match-to-memory values for the guilty suspect and the innocent-suspect and filler distributions represent the range of potential match-to-memory values for the innocent suspect and fillers. To the extent that witnesses can discriminate between the guilty suspect and innocent persons, the guilty-suspect distribution will shift to the right of the filler and innocent-suspect distributions. The vertical dashed lines represent the decision criteria held by the witness. ID_{H} = ID with high confidence; ID_{M} = ID with medium confidence; ID_{L} = ID with low confidence; R_{L} = reject with low confidence; R_{M} = reject with medium confidence; R_{H} = reject with high confidence. The task of the eyewitness is to determine both if the culprit is present in the lineup and whom that person is. A detailed explanation of how the witness uses the decision criteria is described in text.
A standard 6-person culprit-present lineup is represented by one random draw from the guilty-suspect distribution and five random draws from the filler distribution. A standard 6-person culprit-absent lineup is represented by one random draw from the innocent-suspect distribution and five random draws from the filler distribution. The witness' identification decision and confidence in that decision depends on how these six match-to-memory values correspond to her decision criteria. If at least one lineup member exceeds the ID_L criterion the witness affirmatively identifies the best-matching lineup member. If the match value for the best-matching lineup member exceeds the ID_H criterion, the witness identifies that individual with high confidence, if the match value for the best-matching lineup member exceeds the ID_M criterion but falls short of the ID_H criterion, the witness identifies that individual with medium confidence, and if the best-matching lineup member exceeds the ID_L criterion but falls short of the ID_M criterion, the witness identifies that individual with low confidence. Finally, if the match value for the best-matching lineup member falls between ID_L and R_L the model assumes that the witness makes a "not sure" response (Clark, 2003).

Rejection decisions work in a similar manner. If none of the lineup members exceed the R_H criterion, the witness rejects the lineup with high confidence. If none of the lineup members exceed the R_M criterion, the witness rejects the lineup with medium confidence. And, if none of the lineup members exceed the R_L criterion, the witness rejects with low confidence. Hence, the witness is tasked with using her memory for the culprit to determine whether the culprit is present in the lineup procedure and whom that person is. This is commonly referred to as a compound Signal Detection Task (e.g., Duncan, 2006; Palmer et al., 2010) reflecting the fact that the witness simultaneously
completes both a detection task (is the culprit present?) and an identification task (which lineup member is the culprit?).

*The Investigator's Signal Detection Task*

The investigator completes a somewhat simpler detection task. Unlike the witness, the investigator knows which lineup members are the known-innocent fillers and which lineup member is the suspect.\(^2\) Because the investigator knows the identity of the suspect, her detection task conforms to a 2 (decision: arrest, release) x 2 (culprit: present, absent) confusion matrix. Otherwise, put, investigators are not tasked with figuring out who the suspect is (they have this information); investigators are tasked with determining whether the suspect is the person the witness saw commit the crime (a detection task).

A Signal Detection model of the investigator's detection task is illustrated in Figure 2. The model represents the eyewitness evidence that the suspect is guilty and the guilt values for innocent suspects and guilty suspects are represented by two separate Gaussian distributions. To the extent that the investigators are able to use the eyewitness evidence to discriminate between guilty and innocent suspects, the guilty-suspect distribution will be shifted to the right of the innocent-suspect distribution. This makes sense because a guilty suspect actually committed the crime in question and so when the suspect is guilty, the evidence of guilt should tend to be stronger compared to when the suspect is innocent.

---

\(^2\) In fact, the officer who administers the lineup procedure should not know which lineup member is the suspect; that is, lineups should be conducted in a double-blind manner (e.g., Greathouse & Kovera, 2009). But, even though the administrating officer does not know the identity of the suspect, the lead investigator on the case and other members of the department do possess this information.
Figure 2. The guilty-suspect distribution represents the range of potential evidence-of-guilt values for the guilty suspect and the innocent-suspect distribution represents the range of potential evidence-of-guilt values for the innocent suspect. To the extent that the investigator and her lineup procedure can discriminate between guilty suspects and innocent suspects, the guilty-suspect distribution will shift to the right of the innocent-suspect distribution. Investigator discriminability is affected by a combination of the witness’ memory and the quality of the lineup procedure. The vertical dashed lines represent the investigator's criteria for arresting the suspect: ID_H = ID with high confidence; ID_M = ID with medium confidence; ID_L = ID with low confidence; R_H = reject with high confidence; R_M = reject with medium confidence; R_L = reject with low confidence; F_H = filler pick with high confidence; F_M = filler pick with medium confidence; F_L = filler pick with low confidence. Unlike the witness, the investigator already knows who the fillers are and who the suspect is. Indeed, the investigator includes fillers in the lineup as part of her "experimental" design. So the investigator is not tasked with sorting between fillers and suspects. The investigator is tasked only with sorting between guilty suspects and innocent suspects and makes a determination about the suspect's guilt based on the witness’ decision. Suspect identifications are diagnostic of guilt and filler picks and rejections are diagnostic of innocence. The diagnostic value of the witness’ decision increases with the witness’ confidence.

Noticeably, Figure 2 does not include any filler distributions. This makes sense because the investigators know who the fillers are and who the suspect is and so they are not tasked with discriminating between the suspect and fillers. Rather, investigators are tasked only with sorting between guilty suspects and innocent suspects. But fillers still play an important role in the investigator's Signal Detection task. For the investigator's task, filler identifications operate as decision criteria as do suspect identifications and
lineup rejections. Indeed, the investigator makes a decision to arrest or release the suspect based on the decision the eyewitness makes from the identification procedure. In other words, the investigator uses eyewitness responses as different decision criteria (the vertical dashed lines in Figure 2). As noted above, suspect identifications provide evidence of guilt and both filler picks and rejections provide evidence of innocence. Moreover, high-confidence decisions provide stronger evidence of guilt and innocence, respectively, than do decisions made with lower levels of confidence. An investigator with a stringent decision criterion might only make an arrest after a high-confidence suspect identification. An investigator with a more lenient decision criterion might make an arrest even after a low-confidence suspect identification. In theory, an officer could even make an arrest after a filler pick or a rejection. For example, if the prior evidence that a suspect was guilty was particularly strong, an officer might arrest a suspect even after the witness picked a filler out of the lineup or rejected the lineup outright. This might occur for instance if another witness had already identified the suspect with high confidence.

Investigator discriminability is broadly influenced by two factors: (1) the quality of the witness' memory and (2) the quality of the identification procedure. As the quality of either memory or the identification procedure increases, the task of sorting guilty and innocent suspects becomes easier. But, there is some interesting nuance here that is not entirely self-evident. Just as a researcher sets up the protocol for her own experiment, the investigator sets up the protocol for her lineup. How that lineup is designed influences the quality of the witness' decision and, potentially, the witness' memory performance. Hence, the relationship between investigator performance and witness performance is
bidirectional. The procedure that the investigator chooses can affect the performance of
the witness, and of course, the witness’ performance ultimately affects the ability of the
investigator to discriminate between guilty and innocent suspects.

**ROC Analysis of Eyewitness Lineup Procedures**

The distinction between the eyewitness task and the investigator task is important
both for our understanding of how lineups work and for our understanding of how we
ought to analyze lineup data. Ever since Wixted and Mickes (2012, see also Mickes,
Flowe, & Wixted, 2012) introduced the idea of using ROC analysis to analyze data from
eyewitness lineups, it has become, arguably, the most common approach for analyzing
eyewitness lineup data. But, a fundamental misconception about the objective of a lineup
procedure led to the precarious development of comparing *partial* ROC curves. Indeed,
the implicit assumption in past ROC analyses of lineup data is that the witness is the
radiologist and the lineup procedure offers a test of the witness' memory. As we have
explained above, a lineup is not a test of the witness' memory but a test of the
investigator's hypothesis that the suspect is guilty. To appreciate this distinction more
fully consider two contrasting examples recently provided by Wells et al. (2015). Wells
et al. (2015) asked the reader to consider a situation in which an omniscient God appears
immediately before an identification procedure and tells the investigator that the witness' memory is 100% reliable. Does the investigator still want to do the identification
procedure? Absolutely! The investigator still wants to do the identification procedure
because the purpose of the procedure was to test the hypothesis that the suspect is the
culprit and the investigator still does not have any information pertaining to that
hypothesis. Now suppose that on a later investigation, the omniscient God appears again
and tells the investigator that the suspect is, in fact, the culprit. Does the investigator still
want to proceed with the identification procedure? No, because the investigator already
knows that the suspect is guilty and so there is no reason to carry out the lineup procedure
(Wells et al., 2015).

If one falls into the trap of seeing the lineup as a test of the witness' memory and not a test of the suspect's guilt, it is difficult to envision how one could possibly create a full ROC curve based on eyewitness lineup data. Indeed, ROC analysis is designed for dealing with binary-classification tasks. A binary classification task is one in which there are two possible states of the world (e.g., a malignant tumour is present or absent) and two decision outcomes (e.g., an affirmative decision or a negative decision). An eyewitness lineup also has two states of the world (e.g., the culprit is either present in the lineup or absent from the lineup), but the lineup has three possible decision outcomes (suspect pick, filler pick, rejection). Hence, the witness' memory task does not fit the design required to generate an ROC curve. Accordingly, Wixted and Mickes (2012; see also Mickes et al., 2012) recommended that researchers construct only a partial ROC curve that includes only suspect identifications and omits both filler identifications and rejections.

But suspect identifications are not the only eyewitness behaviour that bears on the likely guilt of the suspect. Filler picks and rejections also inform on the likely guilt of the suspect. Each of these three eyewitness behaviors and the witness' associated level of confidence is relevant to the investigator's task of determining whether the suspect is the person the witness saw commit the crime. Accordingly, each of these behaviors should be reflected in any analysis that attempts to discriminate between guilty and innocent
suspects. Moreover, because the investigator's task is a binary classification task (two states of the world: suspect is guilty or suspect is innocent; two decisions: arrest, release), there is no reason not to use a full ROC curve when attempting to discriminate between guilty and innocent suspects.

**The Underlying Logic of ROC Analysis**

For the purpose of explaining the underlying logic of ROC analysis, we consider the example of a one-person showup identification procedure. A showup is a procedure in which law enforcement personnel present a lone suspect to an eyewitness for an identification attempt (e.g., Smith, Wells, Lindsay, & Myerson, 2018). The witness is tasked with indicating whether or not the suspect in the showup is the person she saw commit the crime (the culprit) or an innocent suspect. After making her identification decision, the witness is then asked to qualify this decision with a confidence statement. Hence, a showup identification procedure conforms to a standard signal-detection rating task (Macmillan & Creelman, 2005). For simplicity, we assume that the witness was provided with three confidence options: high, medium, or low.

Figure 3 shows the non-parametric ROC curve for this hypothetical showup procedure. Panel A shows the non-cumulative and cumulative proportions of affirmative-identification (IDs) and negative-identification (rejections) decisions made with high, medium, and low levels of confidence, respectively. Panel B plots the cumulative proportions of culprit-present and culprit-absent decisions in the ROC space. The

---

3 The non-parametric ROC curve is sometimes referred to as the "empirical" ROC curve in the eyewitness literature and parametric ROC curves are sometimes referred to as "theoretical". We find this language misleading. The suggestion is that "empirical" or non-parametric ROC curves permit inference about the applied performance of a lineup procedure and "theoretical" ROC curves permit inference about the memory performance of a lineup procedure. This is absurd. Neither non-parametric nor parametric measures constrain researchers to only making applied or theoretical inferences. Non-parametric simply means that the test does not make detailed distributional assumptions about the data and parametric implies that the test does make detailed distributional assumptions about the data.
leftmost point in Panel B represents the origin of the ROC curve \((y = 0, x = 0)\). The next point in the ROC space plots the proportion of high-confidence affirmative IDs from the culprit-present procedure (on the y-axis) against the proportion of high-confidence affirmative IDs from the culprit-absent procedure (on the x-axis). The next point in the ROC space represents the proportion of affirmative IDs made with \textit{at least} moderate confidence (i.e., the proportion of affirmative IDs made with moderate levels of confidence plus the proportion of affirmative IDs made with high levels of confidence). In other words, the identification decisions cumulate. One continues plotting points in this manner until 100\% of responses from the culprit-present condition and 100\% of responses from the culprit-absent condition are reflected in a single point, the rightmost point on the ROC curve. Once all of the points are plotted in the ROC space, adjacent points are connected with straight lines to produce a non-parametric ROC curve.

<table>
<thead>
<tr>
<th>Choice and Confidence</th>
<th>Non-cumulative</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Culprit Present</td>
<td>Culprit Absent</td>
</tr>
<tr>
<td>Affirmative High</td>
<td>46%</td>
<td>9%</td>
</tr>
<tr>
<td>Affirmative Medium</td>
<td>22%</td>
<td>13%</td>
</tr>
<tr>
<td>Affirmative Low</td>
<td>17%</td>
<td>19%</td>
</tr>
<tr>
<td>Negative Low</td>
<td>9%</td>
<td>21%</td>
</tr>
<tr>
<td>Negative Medium</td>
<td>4%</td>
<td>16%</td>
</tr>
<tr>
<td>Negative High</td>
<td>2%</td>
<td>22%</td>
</tr>
</tbody>
</table>

\textit{Figure 3.} Panel A shows both the non-cumulative and cumulative culprit-present and culprit-absent identification decisions as a function of eyewitness confidence and identification decision. Panel B plots the cumulative identification rates in an ROC curve.
One potential point of confusion here is why negative identification decisions are treated as culprit identifications (for the culprit-present condition) and innocent-suspect identifications (for the culprit-absent condition) in the ROC space. The logic here is that the rating-scale is a behavioral manifestation of the degree to which the suspect matches the witness' memory for the culprit. When a witness makes an affirmative identification decision with high confidence, what the witness is saying is that the suspect provides a very strong match to his/her memory for the culprit. Likewise, when a witness makes a rejection decision with high confidence, what the witness is saying is that the suspect provides a very weak match to his/her memory for the culprit. When a witness' behaviour indicates that the suspect provides a strong match-to-memory (as a high-confidence suspect identification does), this should be taken as relatively strong evidence that the suspect is guilty. Conversely, when a witness' behaviour indicates that the suspect provides a weak match-to-memory (as a high confidence rejection does), this should be taken as relatively strong evidence that the suspect is not guilty. When thinking of these behaviors in terms of how much evidence they provide that the suspect is guilty, the ROC curve is not blending together affirmative and negative decisions, but rather, ordering decisions in terms of how much evidence they provide of the suspects' guilt.

When an ROC curve runs through all possible decision criteria and covers the full range of the X-axis, it provides a threshold-free estimate of discriminability. What this means is that the ROC curve provides an index of performance that is independent of any particular decision threshold or false-positive rate. This becomes important when comparing identification procedures with different false positive rates.
The closer the ROC curve bows to the upper left corner of the ROC space, the better the eyewitness identification procedure is able to discriminate between guilty and innocent suspects. Likewise, when comparing two ROC curves, whichever procedure bows closer to the upper left corner is the procedure with superior diagnostic value (i.e., the procedure that produces a better trade-off between culprit IDs and innocent-suspect IDs). This performance is typically quantified by calculating the area under the ROC curve (AUC). For non-parametric ROC curves, the trapezoidal rule is used to find the AUC. The AUC is equal to:

\[ AUC = \sum_{i=1}^{c-1} \left( \frac{c_{i+1} + c_i}{2} \right) \Delta x \]

Here, \( c_i \) refers to a given cutpoint or confidence bin and \( \Delta x \) refers to the distance between two adjacent cutpoints on the x-axis (the difference between two innocent-suspect ID rates). Hence, we break the ROC curve into \( c - 1 \) trapezoids (6 – 1 = 5 for the example in Figure 3) and for each trapezoid, we multiply the average height (the average culprit ID rate) by the base (the difference in innocent-suspect ID rates between the two cutpoints). Finally, we sum the areas of all trapezoids to find the total AUC. Plugging the cumulative culprit-present and culprit-absent rates from Figure 3 into this equation, we find that the AUC is equal to:

\[ AUC = \left( \frac{46+0}{2} \cdot (0.09 - 0) \right) + \left( \frac{68+.46}{2} \cdot (.22 - .09) \right) + \cdots + \left( \frac{100+.98}{2} \cdot (1.00 - .78) \right) \]

\[ AUC = .80 \]

Calculation of the AUC should make it readily apparent why the procedure producing a larger AUC should be preferred. The procedure with the larger AUC is the procedure that, on average, leads to a higher culprit-identification rate. Because the AUC is
calculated over the full range of potential innocent-suspect identification rates, we have effectively cancelled out or controlled for the innocent-suspect identification rate and therefore, we should prefer whichever procedure produces the higher AUC (i.e., the higher average culprit ID rate). In this sense, we have obtained a threshold-free estimate of performance.

Before moving any further, one final point is worth noting here. For the 2 (culprit: present, absent) x 2 (witness decision: affirmative, negative) showup identification procedure, the witness and the investigator share the same task. Both the witness and the investigator are tasked only with determining whether the suspect is the culprit from the crime or an innocent suspect. The investigator might use different criteria than the witness (e.g., the witness identifies the suspect with low confidence, but the investigator decides this is insufficient evidence for arrest), but the discrimination task is one and the same. It is only when the investigator introduces fillers to create a lineup procedure that the investigator's task becomes appreciably different from the witness' task.

The Logic of ROC Analysis Falls Apart with Partial ROC Curves

Contrast the (full) ROC curve in Figure 3 that is typical of many applied sciences (e.g., Hanson, Lloyd, Helmus, & Thornton, 2012; McPherson, Steel, & Dixon, 2000; Pisano et al., 2005) with the partial ROC curves that have become commonplace for the eyewitness lineup literature (see Figure 4). Figure 4 displays partial high-similarity lineup and low-similarity lineup ROC curves from Colloff et al. (2016). The logic of surrounding the suspect with high-similarity fillers is that if the suspect is innocent, high-similarity fillers prevent that person from standing out and lower the risk of mistaken identification. To the extent that fillers are high in similarity to the suspect, the
probability of an innocent suspect identification is decreased because many mistaken identifications that could potentially land on the innocent suspect instead land on the known innocent-fillers (e.g., Colloff et al., 2016; Smith et al., 2018). Unfortunately, high-similarity fillers also lead to lower rates of culprit identification than do low-similarity fillers (Clark, 2012; Fitzgerald, Price, Oriet, & Charman, 2013). Hence, the difference between high-similarity and low-similarity lineups is defined by a trade-off: high-similarity lineups lead to both fewer innocent-suspect identifications and fewer culprit identifications than do low-similarity lineups. When the difference between two procedures is defined by a trade-off it is difficult to champion one over the other, because both come with a benefit and a cost. Unlike other analytic techniques, (full) ROC analysis was designed specifically to deal with these trade-off problems.

Figure 4. Partial ROC for high-similarity lineups and low-similarity lineups from Colloff et al. (2016).
Notice that the ROC curves in Figure 4 look far different than did the hypothetical ROC curve depicted in Figure 3. Indeed, the lineup ROC curves in Figure 4 cover only a small portion of the X-axis. This happens because fillers siphon many identifications away from innocent suspects and thus, the maximum innocent-suspect identification rate will be far less than 1.00. Because these curves span only a small portion of the X-axis, lineup researchers calculate a partial AUC (pAUC) rather than a full AUC as is typical in other applied sciences. When two curves span the same region of the X-axis, choosing a pAUC region over which to compare the two procedures is straightforward: the curves are compared over the X-axis region from .00 to the point at which the ROC curves terminate (the innocent-suspect identification rate that is achieved after cumulating across all levels of confidence). But, comparing pAUC values becomes problematic when ROC curves span different regions of the X-axis (i.e., have different cumulative innocent-suspect identification rates) (Lampinen et al., 2019; Smith et al., 2019). Whereas the low-similarity lineup in Figure 4 covers potential innocent-suspect identification rates from .00 to .36, the high-similarity lineup only covers the range of .00 to .09. Here we encounter a problem that is not encountered when scientists compare full ROC curves: how do we compare lineup ROCs that span different regions of the X-axis? There appears to be three possibilities (Smith, Smalarz, & Jalava, accepted):

1. Compare the two curves over each curve's full range.
2. Compare the two curves over the largest curve's range.
3. Compare the two curves over the shortest curve's range.

Unfortunately, none of these practices is without problems. First, Option 1 is clearly inappropriate. Holding all else constant, the procedure that covers a wider range on the
X-axis would have a larger pAUC simply because it has a wider base. But, a wider base corresponds to a higher innocent-suspect identification rate, so Option 1 would reward the procedure that resulted in more innocent-suspect identifications, when, if anything, a procedure should be penalized for resulting in more innocent-suspect identifications. Option 2 is also problematic as it would require making parametric assumptions and extrapolating the shorter curve well beyond the observed data for that procedure. But, we do not really know where that curve will project to and this extrapolation approach is really a guessing game (Collof et al., 2016). Even more problematic than that, the innocent-suspect identification rate for the high-similarity lineup cannot possibly exceed \(\frac{1.00}{6\text{ lineup members}}\), so Option 2 would also require extrapolating far beyond the theoretically permissible values for that procedure (to an innocent-suspect identification rate of .36) (Smith et al., 2019; Smith et al., accepted).

Option 3 has become the default for researchers using partial ROC analysis to compare lineup procedures with different innocent-suspect identification rates. Collof et al. (2016) for example compared their high- and low-similarity lineups over a range of innocent-suspect identification rates from .00 to .09 (the region covered by their high-similarity lineup). The problem with Option 3 is that it requires the researchers to discard large portions of suspect identifications from the procedure with the higher innocent-suspect identification rate.\(^4\) Indeed, the portion of the low-similarity lineup that exceeds an innocent-suspect identification rate of .09 was excluded from the pAUC analysis. The result is that the researchers actually compared all suspect identifications from the high-

\(^4\) This is on top of the fact that partial ROC curves always discard the large portions of witness-participants who identify fillers or reject the lineup procedure.
similarity lineup to only those suspect identifications that were made with 100% confidence from the low-similarity lineup. The researchers concluded that high-similarity lineups were objectively superior to low-similarity lineups (Colloff et al., 2016). This conclusion is overgeneralized. At best, the researchers could conclude that the diagnostic value of suspect identifications from the high-similarity lineup was superior to the diagnostic value that the low-similarity lineup can achieve for suspect identifications made with 100% confidence (Smith et al., 2019; Smith et al., accepted). The more general conclusion that high-similarity lineups are superior to low-similarity lineups is not warranted because the pAUC analysis discards the majority of the data from the low-similarity lineup. The low-similarity lineup is not permitted to realize the advantage it has relative to a high-similarity lineup: a higher culprit identification rate.

In fact, using a classic utility analysis, Smith et al. (2019; see also Lampinen et al., 2019) demonstrated that under certain assumptions about base rates and about the relative costs of missed culprit identifications and innocent-suspect identifications, the low-similarity lineup is actually superior to the high-similarity lineup. We revisit the utility approach to comparing lineups later in this manuscript, but suffice to say the pAUC approach to comparing lineups is inappropriate. We now examine more generally why the pAUC approach fails to identify the diagnostically superior procedure.

Why Comparison of Partial ROC Curves is Problematic

The goal of ROC analysis is to provide a threshold-free measure of performance. In other words, the goal is to identify which procedure better discriminates between guilty and innocent-suspects, independent of any specific false-alarm rate. The full ROC curve achieves this by plotting a curve that covers the full range of the X-axis (see Figure
3). By plotting the cumulative culprit-identification rate over the full range of the X-axis (i.e., the full range of potential innocent-suspect identification rates), one has controlled for or partialed out the innocent-suspect identification rate. Once we have controlled for the innocent-suspect identification rate, we should prefer whichever procedure, on average, gives us a higher culprit identification rate. This is what the AUC measures: the average culprit-identification rate that a procedure achieves for the full range of potential innocent-suspect identification rates. Hence, to prefer the procedure with the larger AUC is simply to prefer the procedure with the higher culprit identification rate after controlling for the innocent-suspect identification rate.

Because partial ROC curves do not span the full range of the X-axis (the full range of potential false alarm rates), they do not provide a threshold-free index of performance. In other words, which procedure the partial ROC curve favors is confounded with the false alarm rate. As we demonstrate below, pAUC analysis is biased towards favoring the procedure with the lower innocent-suspect identification rate.

Figure 5 shows the diagnostic values of two testing procedures, one relatively conservative and the other relatively liberal. On the left side of the figure, we plot the distributions of true-positive and false-positive rates across a range of decision criteria (or thresholds). Zero represents the true-positive rate that is achieved at a false-positive rate of 0 and c4 through c1 represent additional decision criteria in descending order of stringency. The conservative procedure produced true-positive and false-positive rates of 28% and 9%, respectively, at the c4 threshold. All remaining responses were made at the c1 threshold. The liberal procedure produced true-positive and false-positive rates of 57% and 36%, respectively, at the c3 threshold. Again, all remaining responses were made at
the c1 threshold. Note that we could have used more complex distributions, but decided to use only two decision criteria for each procedure in order to keep this example as simple as possible. Also note that the c4 values for the conservative procedure correspond to the cumulative culprit and innocent-suspect identification rates for Colloff et al.'s (2016) high-similarity lineup and the c3 values for the liberal procedure correspond the cumulative culprit and innocent-suspect identification rates for Colloff et al.'s (2016) low-similarity lineup.
Figure 5. Comparing the AUC values for two procedures that are equivalent in diagnostic value. The top row displays performance for a relatively conservative procedure and the bottom row displays performance for a relatively conservative procedure. Although how the responses are distributed differs between these two procedures, the procedures are equivalent in terms of diagnostic accuracy (i.e., the trade-off between true positives and false positives).

The right side of Figure 5 shows the ROC curves that are produced by the distributions on the left. The conservative procedure results in an AUC of .60 and the liberal procedure results in an AUC of .61. For practical purposes, these two procedures have equal diagnostic value. Given that these procedures represent the cumulative culprit
and innocent-suspect identification rates from Colloff et al.'s (2016) high-similarity and low-similarity lineups, how is it that Colloff et al. (2016) came to conclude that the high-similarity lineup has superior diagnostic value to that of the low-similarity lineup?

**pAUC Analysis is Biased in Favor of More Conservative Procedures.** Figure 6 shows the partial ROC comparison that results from the data in Figure 5. Note that these ROC curves closely resemble the high- and low-similarity ROC curves from Colloff et al. (2016) that are displayed in Figure 4 of the present manuscript. Following common practice for comparing partial lineup ROC curves, we calculated pAUC values for the region of the X-axis spanned by the more conservative procedure (.00 -.09). Over this restricted region of the X-axis, the conservative procedure produced a pAUC (.013) that was almost twice the size of the pAUC achieved by the liberal procedure (.006). This happens in spite of the fact that when we compared the full AUC values for these two procedures, there was no difference in AUC values.
Figure 6. Comparison of pAUC values for the hypothetical conservative and liberal ROC curves from Figure 5. Even though the full AUC shows that the two procedures are equivalent in diagnostic value (see Figure 5), the pAUC analysis shows a strong preference for the more conservative procedure.

That this pAUC approach is biased in favor of the more conservative procedure is a product of simple geometry. Holding discriminability equal, the more conservative of two procedures will produce a steeper slope from .00 to the cumulative innocent-suspect identification rate (the terminal points in Figure 6). Accordingly, if we compare two procedures over this limited range of the X-axis, the pAUC will be biased in favor of the more conservative procedure. The reason the full AUC does not carry the same bias is because the more liberal procedure will have a steeper slope from its cumulative
innocent-suspect identification rate (the terminal point in Figure 6) to 1.00 than will the more conservative procedure. Hence, the full AUC has a way of balancing things out that is not present when comparing partial lineup ROC curves.

**How to Build Full ROC Curves for Eyewitness Lineups**

As noted previously, the inappropriate use of partial ROC curves resulted from a fundamental misunderstanding of how lineups work. The purpose of a lineup procedure is to test the investigator's hypothesis that the suspect committed some crime in question. The witness' identification decision and her confidence in that identification decision provide evidence of the suspect's guilt that the investigator uses to make an arrest decision. Hence, it is the investigator and not the witness who is acting in the role of the radiologist. The witness acts in the role of the X-ray. Like a high-resolution X-ray, a witness with a strong memory provides the investigator with more information than does a witness with a weak memory. With this in mind, the full ROC curve for a lineup procedure should conform to the investigator's Signal Detection task and not to the eyewitness' Signal Detection task. From this realization, a method for creating a full eyewitness lineup ROC curve becomes evident.

*A re-examination of the high- vs. low-similarity comparison from Colloff et al. (2016)*

The police investigator's task is to use the witness' decision and her confidence in that decision to make an inference about whether the suspect is guilty or innocent. As demonstrated in Figure 2, investigators have control over how much evidence of guilt they require to arrest the suspect. Some might only make an arrest after a high-confidence suspect identification, but other investigators (or in different contexts) might make an
arrest after even a low-confidence rejection. Hence, the witness decisions and associated confidence levels reflect different operating points on the investigator's ROC curve.

In creating a full lineup ROC curve, we encounter a difficulty that we did not encounter when we plotted the hypothetical (full) showup ROC curve in Figure 3. That is, how do we determine which eyewitness responses provide the strongest evidence of guilt (and weakest evidence of innocence) and which eyewitness responses provide the weakest evidence of guilt (and strongest evidence of innocence)? In Figure 3, we assumed that suspect identifications provide stronger evidence of guilt (and weaker evidence of innocence) than do rejections and that decisions made with higher levels of confidence are more informative (of guilt or innocence) than are decisions made with lower levels of confidence. With lineups, it would also be safe to assume that suspect identifications are more diagnostic of guilt than are filler picks or rejections (e.g., Wells et al., 2015). But, how do we order filler picks and rejections? Rather than trying to determine whether suspect picks or filler picks are more diagnostic of innocence, a priori, we argue that it would be more appropriate to calculate the likelihood of guilt (i.e., the diagnosticity ratio) for each operating point on the lineup ROC curve and to sort those operating points based on their relative likelihood of guilt.

Historically, eyewitness researchers used the diagnosticity ratio to compare the relative performance of different lineup procedures. The diagnosticity ratio reflects the likelihood that a suspect is guilty given that she was identified and is calculated by dividing the culprit identification rate by the innocent-suspect identification rate (Wells & Lindsay, 1980). Some have argued that the diagnosticity ratio is primarily concerned with measuring the conservativism of a lineup procedure (or how well it does at protecting
innocent suspects) rather than discriminability (or how good of a trade-off the procedure produces between culprit and innocent-suspect identifications) (e.g., Wixted & Mickes, 2014). Indeed, the diagnosticity ratio is affected both by discriminability and by how conservative that procedure is. If conservativism is held constant, the procedure with greater discriminability will produce a higher diagnosticity ratio. Likewise, if discriminability is held constant, the more conservative procedure will produce a higher diagnosticity ratio. But, the discriminability of a procedure is defined by its ROC curve; hence every operating point on the same ROC curve shares the same discriminability (Wixted & Mickes, 2012). Within a procedure, different operating points reflect different levels of conservativism rather than differences in discriminability. Accordingly, we calculated the diagnosticity ratio for each point on a given ROC curve and then sorted these points based on their relative evidence of guilt.

Table 1 shows the non-cumulative hit rates and false-alarm rates for the high-similarity and low-similarity lineups from Colloff et al. (2016). In order to make this example tractable, we collapsed adjacent confidence bins to decrease the number of operating points from 33 operating points to nine operating points. It is evident from Table 1 that identification decisions and associated levels of confidence were correlated with the likelihood of guilt. For example, suspect identifications were more diagnostic of guilt than were either filler picks or rejections. Likewise, high-confidence decisions tended to be more informative (more discrepant from 1.00) than low-confidence decisions. Yet, there were some anomalies. For example, for the high-similarity lineups, filler picks made with 0% - 60% confidence were more diagnostic of innocence (or less diagnostic of guilt) than were filler picks made with higher levels of confidence.
Likewise, rejections made with 70% - 80% confidence were more diagnostic of innocence than were rejections made with 90% - 100% confidence. All this is to say that confidence is not perfectly associated with the likelihood of guilt.

Table 1: Non-cumulative true-positive and false positive rates as a function of the witness' identification decision and associated level of confidence

<table>
<thead>
<tr>
<th>Evidence</th>
<th>High Similarity</th>
<th>Low Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>S 90-100</td>
<td>.065 .008 8.18</td>
<td>S 90-100 .240 .120 2.00</td>
</tr>
<tr>
<td>S 70-80</td>
<td>.063 .019 3.23</td>
<td>S 70-80 .140 .073 1.92</td>
</tr>
<tr>
<td>S 0-60</td>
<td>.158 .061 2.60</td>
<td>S 0-60 .140 .111 1.13</td>
</tr>
<tr>
<td>R 0-60</td>
<td>.201 .245 0.85</td>
<td>F 90-100 .022 .022 1.00</td>
</tr>
<tr>
<td>F 70-80</td>
<td>.083 .097 0.85</td>
<td>F 70-80 .124 .146 0.85</td>
</tr>
<tr>
<td>F 90-100</td>
<td>.033 .040 0.83</td>
<td>F 90-100 .040 .048 0.82</td>
</tr>
<tr>
<td>F 0-60</td>
<td>.231 .305 0.76</td>
<td>R 0-60 .132 .211 0.62</td>
</tr>
<tr>
<td>R 90-100</td>
<td>.072 .099 0.73</td>
<td>R 70-80 .059 .105 0.57</td>
</tr>
<tr>
<td>R 70-80</td>
<td>.086 .126 0.68</td>
<td>R 90-100 .057 .110 0.52</td>
</tr>
</tbody>
</table>

Note. Evidence = witness decision and associated level of confidence HR = hit rate or culprit arrest rate; FAR = false alarm rate or innocent-suspect arrest rate; DR = diagnosticity ratio of guilt or the likelihood that the suspect is guilty given the witness response; S = suspect; F = filler, and R = rejection. The numeric values in the evidence column refer to the confidence bin.

The two ROC curves depicted in Figure 7 paint a very different picture than do the partial ROC curves depicted in Figure 4. The partial ROC curves depicted in Figure 4 appear to show that the high-similarity lineup is superior to the low-similarity lineup. Figure 7 makes it clear that the high-similarity lineup does not produce a better trade-off between culprit and innocent-suspect identifications than does the low similarity lineup. The high-similarity lineup produced an AUC of .62 and the low-similarity lineup produced an AUC of .63. In fact, the difference between the partial ROC curves depicted in Figure 4 and the full ROC curves depicted in Figure 7 almost perfectly resembles the difference between the hypothetical full ROC curves and partial ROC curves depicted in Figure 5 and Figure 6, respectively. In other words, the difference between the high-similarity and low-similarity lineups appears to be one that is better characterized by a change in conservativism rather than by a change in discriminability (cf. Colloff et al.,
2016; cf. Smith et al., 2018). One might still prefer the high-similarity lineup over the low-similarity lineup, but only to the extent that one is willing to make detailed assumptions about the underlying base rates of culprit presence and the relative costs of missed culprit identifications and innocent suspect identifications (Lampinen et al., 2019; Smith et al., 2019). We revisit these assumptions prior to our conclusion.

**Figure 7.** Full ROC comparison of the high-similarity and low-similarity lineups from Colloff et al. (2016). As is evident from the ROC curves, the difference between high-similarity lineups and low-similarity lineups is better characterized by a change in conservativism than by a change in discriminability. The high-similarity lineup has an AUC of .62 and the low-similarity lineup has an AUC of .63.

**Comparing Full Lineup ROC Curves for Strong and Weak Memory Conditions**

In order to contrast the difference between a change in conservativism produced by manipulations of filler similarity with a change in discriminability, we also re-examined a manipulation of memory strength. We re-examined a recent manipulation of encoding conditions from Smith, Wilford, Quigley-McBride, and Wells (2019,
Smith et al. (2019) randomly assigned participant-witnesses to watch either a clear or degraded version of a simulated crime video. Afterwards, participants were randomly assigned to view either a culprit-present or a culprit-absent lineup procedure. Lineup performance for the clear and degraded viewing conditions is depicted in Table 2. As with the Colloff et al. (2016) data, we created three confidence categories: low confidence (0% - 60%), medium confidence (70% - 80%), and high confidence (90% - 100%). We then sorted eyewitness identification decisions and associated levels of confidence based on their associated likelihoods of guilt.

Table 2: Non-cumulative true-positive and false positive rates as a function of the witness' identification decision and associated level of confidence

<table>
<thead>
<tr>
<th></th>
<th>Clear</th>
<th>Degraded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence</td>
<td>HR</td>
<td>FAR</td>
</tr>
<tr>
<td>S High</td>
<td>.383</td>
<td>.000</td>
</tr>
<tr>
<td>S Medium</td>
<td>.250</td>
<td>.013</td>
</tr>
<tr>
<td>S Low</td>
<td>.150</td>
<td>.029</td>
</tr>
<tr>
<td>R Low</td>
<td>.167</td>
<td>.306</td>
</tr>
<tr>
<td>F Low</td>
<td>.017</td>
<td>.116</td>
</tr>
<tr>
<td>R Medium</td>
<td>.017</td>
<td>.242</td>
</tr>
<tr>
<td>R High</td>
<td>.017</td>
<td>.242</td>
</tr>
<tr>
<td>F Medium</td>
<td>.000</td>
<td>.052</td>
</tr>
<tr>
<td>F High</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note. Evidence = witness decision and associated level of confidence HR = hit rate or culprit arrest rate; FAR = false alarm rate or innocent-suspect arrest rate; DR = diagnosticity ratio of guilt or the likelihood that the suspect is guilty given the witness response; S = suspect; F = filler, and R = rejection. High = 90% - 100% confidence; Medium = 70% - 80%; Low = 0% - 60%. +Inf = positive infinity*

As was the case with the Colloff et al. (2016) data, we can see that the likelihood of guilt is strongly associated with the witness' decision and associated confidence level. In fact, under clear viewing conditions, this relationship is perfectly ordered. This is not unexpected. To the extent that witnesses have strong memories, we should expect eyewitness decisions and associated levels of confidence to be more informative. Hence, suspect identifications should become far more likely to occur when the culprit is present compared to when the culprit is absent and filler identifications and rejections should
become far more likely then the culprit is absent compared to when the culprit is present. Moreover, while high-confidence suspect identifications should occur relatively frequently when the culprit is present, they should be extremely infrequent when the culprit is absent. Likewise, moderate and high confidence rejections and filler picks should be relatively likely when the culprit is absent, but very unlikely when the culprit is present (because there is such a strong signal emanating from the culprit, that those who pick a filler or reject when the culprit is present would not be able to do so with high confidence). To the contrary, when memory is weak, rejections and filler picks should increase even when the culprit is present, making them less diagnostic of innocence. Likewise, when memory is weak, confidence in all identifications should decrease. Hence, witness behaviors and associated levels of confidence should become less diagnostic across the board and separation among the different witness behaviors should decrease. All this is to say that when memory strength is decreased, the investigator's task becomes more difficult.

Figure 8 shows the ROC curves for the strong (clear view) and weak (degraded view) memory conditions from Smith et al. (2019, Experiment 1). The strong memory condition dominates the weak memory condition across the entire range of false-positive rates. Indeed, the strong-memory condition had an AUC of .93 and the weak-memory condition had an AUC of only .61. What we might infer from this is that the strong-memory condition is universally superior to the weak-memory condition. What we mean by this is that, no matter the underlying base rates and no matter the relative costs of missed culprit identifications relative to innocent-suspect identifications, the strong memory condition will be superior to the weak memory condition. This is one of the
chief benefits of using full ROC analysis to compare lineup procedures: if one condition dominates the other condition over the full range of false-positive rates, we know that the dominating procedure will always be superior.

Figure 8. An example of two full lineup ROC curves when there is a difference in discriminability. Here, the clear-view lineup ROC curve dominated the degraded-view ROC curve over the full range of potential false alarm rates.

**How Do We Proceed when ROC Curves Crossover?**

When one ROC curve dominates another ROC curve over the full range of potential false-positive rates, this implies that the dominating procedure is universally superior, meaning that there are no foreseeable conditions under which the dominated curve might prove superior. An interesting question concerns what researchers should do when full ROC curves intersect or crossover. We suspect that crossovers will prove common in the eyewitness lineup literature (see Clark, 2012, for example). But, even when ROC curves do crossover, there might be reasons to prefer one lineup procedure to
another lineup procedure. A more conservative procedure might be justified on the grounds that the cost of innocent-suspect identifications exceeds the costs of missed culprit identifications, on the grounds that the base rate of culprit presence is low, or based on a combination of these two assumptions. For example, one could use these assumptions to argue that a high-similarity lineup has better utility than does a low-similarity lineup (Clark, 2012; Lampinen et al., 2019; Smith et al., 2019). But, how does one go about formally quantifying this preference? As far as we can tell there are two possibilities: (1) follow-up a full ROC analysis with a partial ROC analysis or (2) follow-up a full ROC analysis with a formal utility analysis. We review each of these options in turn.

When two ROC curves intersect, this suggests that over some range of potential false-positive rates, Procedure A might be superior to Procedure B, and over some other range of potential false-positive rates, the reverse might be true. In order to test the significance of differences over these partial ranges, instinctively one might consider implementing a series of pAUC analyses. Indeed, one could specify a range of false-positive rates over which to compare the lineup ROC curves and compare the relative pAUC values over that range. But, when one selects a limited range of false-positive rates over which to compare two ROC curves that individual is making a number of assumptions about the underlying base rates and about the relative costs of innocent-suspect identifications and missed culprit identifications that are left implicit in this analysis (Lampinen et al., 2019; Smith et al., 2019). In our view, a better approach is to use an expected utility analysis (Lampinen et al., 2019; Smith et al., 2019; Yang et al., 2019). Unlike pAUC analysis, expected utility analyses bring assumptions about
underlying base rates and the relative costs of missed culprit and innocent-suspect identifications into the light of day (Clark, 2012; Lampinen, 2016). Indeed, utility analyses force researchers to explicitly outline the specific assumptions that are required to prefer one lineup procedure to another.

We believe that it is worthwhile for researchers to start by using a full ROC analysis to compare lineup procedures because if one procedure dominates another procedure over the full range of false-positive rates (from 0.00 to 1.00), this means that the dominating procedure is superior under any and all assumptions about base rates and about the relative costs of missed culprit and innocent-suspect identifications. But, when ROC curves intersect, as was the case in our reanalysis of Colloff et al. (2016; see our Figure 7), we believe that an expected utility analysis has more to offer as a follow-up than does a pAUC analysis. As we have shown elsewhere, to the extent that the costs of innocent-suspect identifications exceed the costs of missed culprit identifications or to the extent that the base rate of culprit presence is low, one should prefer a high-similarity lineup to a low-similarity lineup. But, to the extent that one is unwilling to make those assumptions, there is no reason to prefer the high-similarity lineup over the low-similarity lineup and under some relatively optimistic assumptions one might even prefer a low-similarity lineup to a high-similarity lineup (Lampinen et al., 2019; Smith et al., 2019).

Without some sort of expected utility analysis, we cannot identify the basis for our preference.

**Conclusion**

The present work makes an important distinction that has gone under the radar for more than 40 years: the distinction between witness discriminability and investigator
When it comes to completing an eyewitness lineup procedure, the witness completes a 3 (behaviour: suspect pick, filler pick, rejection) x 2 (culprit: present, absent) task in which she must determine both (1) whether the culprit is present or absent in the lineup procedure (a detection task) and (2) whom that person is. To the contrary, the investigator knows which lineup members are fillers and which lineup member is the suspect. The investigator has access to this information because the investigator designed the "experiment". Accordingly, the only task for the investigator is to determine whether the suspect is the culprit that committed the crime in question. It is the investigator's task that is important for applied purposes as no matter how the witness responds, the investigator is the one who controls the fate of the suspect. Because the investigator's task is a binary classification task, this opens up the possibility to examine eyewitness lineup procedures with full ROC curves rather than with the problematic partial ROC curves that have proliferated the literature to date. The key is to recognize that the witness' decision and expressed level of confidence bears on the likely guilt of the suspect. High-confidence suspect identifications are most diagnostic of guilt and high-confidence rejections and filler picks are least diagnostic of guilt (and most diagnostic of innocence). After putting these behaviors and confidence values in descending order of likely guilt, the full ROC curve is created by plotting these values in the ROC space. This new approach to analyzing eyewitness lineup data will lead to major advancements in our understanding of how lineup procedures work. Contrary to partial ROC analyses, the present work demonstrates that the difference between high-similarity lineups and low-similarity lineups is better characterized by a change in conservativism than by a change in diagnostic value.
References


