Applied Cognitive Psychology, Appl. Cognit. Psychol. **31**: 467–477 (2017)
Published online 3 August 2017 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.3344

ROCs in Eyewitness Identification: Instructions versus Confidence Ratings

LAURA MICKES[1]*, TRAVIS M. SEALE-CARLISLE[1], STACY A. WETMORE[2], SCOTT D. GRONLUND[3], STEVEN E. CLARK[4], CURT A. CARLSON[5], CHARLES A. GOODSELL[6], DAWN WEATHERFORD[7] and JOHN T. WIXTED[8]

[1]*Royal Holloway, University of London, Egham, UK*
[2]*Butler University, Indianapolis, USA*
[3]*University of Oklahoma, Norman, USA*
[4]*University of California, Riverside, Riverside, USA*
[5]*Texas A&M University–Commerce, Commerce, USA*
[6]*Canisius College, Buffalo, USA*
[7]*Texas A&M University–San Antonio, San Antonio, USA*
[8]*University of California, San Diego, San Diego, USA*

Summary: From the perspective of signal detection theory, different lineup instructions may induce different levels of response bias. If so, then collecting correct and false identification rates across different instructional conditions will trace out the receiver operating characteristic (ROC)—the same ROC that, theoretically, could also be traced out from a single instruction condition in which each eyewitness decision is accompanied by a confidence rating. We tested whether the two approaches do in fact yield the same ROC. Participants were assigned to a confidence rating condition or to an instructional biasing condition (liberal, neutral, unbiased, or conservative). After watching a video of a mock crime, participants were presented with instructions followed by a six-person simultaneous photo lineup. The ROCs from both methods were similar, but they were not exactly the same. These findings have potentially important policy implications for how the legal system should go about controlling eyewitness response bias.

Wixted and Mickes (2012) argued that when the goal is to measure how well eyewitnesses can discriminate between innocent and guilty suspects, plotting the receiver operating characteristic (ROC) is more appropriate than computing the diagnosticity ratio. Both approaches take into account the overall correct and false identification (ID) rates (the proportion of participants who correctly identify guilty suspects and incorrectly identify innocent suspects, respectively), but ROC analysis also takes into account the additional correct and false ID rates that can be computed as the willingness to make an ID varies from liberal to conservative (Gronlund, Wixted, & Mickes, 2014). Each pair of correct and false ID rates constitutes one point on the ROC. The more the family of ROC points from a given condition bow up and away from the diagonal line of chance performance, the better able eyewitnesses are to sort innocent and guilty suspects into their correct categories—that is, the better able they are to discriminate innocent from guilty suspects.

Although several different methods can be used to generate ROC data, only one method has been used thus far in the eyewitness ID literature (e.g., Anderson, Carlson, Carlson, & Gronlund, 2014; Carlson & Carlson, 2014; Colloff, Wade, & Strange, 2016; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Flowe, Klatt, & Coloff, 2014; Flowe, Smith, Karoğlu, Onwuegbusi, & Rai, 2015a, 2015b; Humphries & Flowe, 2015; Key et al., 2015; Lampinen, Erickson, Moore, & Hittson, 2014; Mickes, 2015; Neuschatz et al., 2016; Seale-Carlisle & Mickes, 2016; Smith & Flowe, 2015; Wetmore, Neuschatz, Gronlund, Key, & Goodsell, 2015a; Wetmore, Neuschatz, Gronlund, Wooten, Goodsell, & Carlson, 2015b). That method makes use of confidence ratings that participants provide when they make an ID from a lineup (for a tutorial, see Gronlund et al., 2014). The first point on the confidence-based ROC is obtained by computing the correct and false ID rates in the usual way, namely, by counting all suspect IDs regardless of the confidence expressed by the participant. This (most liberal) ROC point is associated with the highest correct and false ID rates for a given condition, and these are the correct and false ID rates that have long been used to compute the diagnosticity ratio (correct ID rate/false ID rate). Additional (more conservative and therefore lower) correct and false ID rates are computed by setting an ever-higher standard on the confidence scale for counting IDs. Thus, for example, the second ROC point is obtained by counting all suspect IDs except those that were made with the lowest level of confidence (i.e., by treating as a non-ID any suspect ID that is acknowledged by the participant to rely on little mnemonic support). The last ROC point is computed by counting only suspect IDs that were made with the highest level of confidence. This (most conservative) ROC point is associated with the lowest correct and false ID rates for a given condition.

An alternative method for generating ROC data—one that does not rely on confidence ratings—uses pre-test instructions to manipulate response bias from liberal to conservative (Macmillan & Creelman, 2005). In the liberal response bias condition, the instructions encourage the participant to make an ID from the lineup, resulting in relatively high correct and false ID rates. In the conservative response bias condition, by contrast, the instructions discourage the participant from making an ID unless a participant is quite certain, resulting in relatively low correct and false ID rates. In a neutral response bias condition, the instructions neither encourage nor discourage the participant from making an ID (resulting in intermediate correct and false ID rates). When the correct and false ID rates from the different biasing

*Correspondence to: Laura Mickes, Royal Holloway, University of London, Egham, UK.
E-mail: laura.mickes@rhul.ac.uk

conditions are plotted against each other on a graph, they make up the instruction-based ROC. As with the confidence-based ROC, the more those points bow up and away from the diagonal line of chance performance, the better able eyewitnesses are to discriminate innocent from guilty suspects.

Confidence-based and instruction-based ROCs have been found to be similar to each other when a list memory procedure is used (e.g., Dubé & Rotello, 2012; Koen & Yonelinas, 2011). In these studies, participants first study a list of items (e.g., words) and are then presented with a recognition test in which they make an old/new recognition decision for each of many targets and lures (items that did or did not appear on the list, respectively). However, the two methods of constructing an ROC have not been compared using an eyewitness ID procedure in which participants first witness a mock crime and are then tested only once (e.g., using a photo lineup). The purpose of the research reported here is to do just that.

The comparison between confidence-based and instruction-based ROCs has potentially important policy implications. For example, standard lineup instructions stipulate that the perpetrator may or may not be in the lineup (neutral response bias). Policymakers in jurisdictions where false ID rates are thought to be unacceptably high might consider changing these standard instructions in such a way as to induce a more conservative response bias. Doing so would reduce the false ID rate but at the potential cost of reducing the correct ID rate as well. Does science have any useful information to provide in helping jurisdictions to make that decision? Science cannot help with the value-based question of whether or not, all else being equal, the loss of correct IDs is worth the reduction in false IDs (Clark, 2012), but it can help to establish whether or not instructions designed to induce more conservative responding has any effect on the ability of eyewitnesses to discriminate innocent from guilty suspects (i.e., whether or not all else is equal). If, for example, a specific set of instructions not only induced more conservative responding but also reduced discriminability, then the best course of action might be to seek an alternative strategy to reduce the false ID rate—one that does not reduce discriminability. Considerations like these explain why a National Academy of Sciences committee recently made the following recommendation: 'The committee thus recommends a rigorous exploration of methods that can lead to more conservative responding (such as witness instructions) but do not compromise discriminability' (p. 118, National Research Council, 2014). The work we report here was conducted in response to that recommendation.

Prior research on lineup instructions

Prior research on the effect of lineup instructions has often compared 'biased' with 'unbiased' instructions (see Steblay, 1997, for a review of this literature). Biased instructions encourage participants to make an ID (i.e., biased instructions encourage a liberal response bias), whereas unbiased instructions neither encourage nor discourage participants to make an ID (i.e., unbiased instructions encourage a more conservative intermediate response bias). These studies did not perform ROC analysis but instead relied on the diagnosticity ratio to compare performance in the two lineup instruction conditions. The diagnosticity ratio is the correct ID rate divided by the false ID rate, and it indicates the likelihood that an identified suspect is actually guilty (i.e., it indicates the trustworthiness of suspect IDs in each instructional condition). The assumption has long been made that the better lineup instruction is the one that results in more trustworthy suspect IDs. For example, Lindsay et al. (1991) advanced the following argument:

> Biased lineup procedures consistently resulted in lower diagnosticity ratios (i.e., lower ratios of correct to false identifications, Wells & Lindsay, 1980) than do unbiased lineups. Higher diagnosticity ratios should result in greater probative value, which leads to strong recommendations that biased lineups be avoided (p. 796).

However, as has been frequently noted in the recent debate over simultaneous versus sequential lineups, as responding becomes more conservative, the diagnosticity ratio naturally increases even if discriminability remains constant (e.g., Gronlund et al., 2014). In fact, the diagnosticity ratio can increase with more conservative responding even if discriminability *decreases*. Thus, prior work on the effect of lineup instructions did not address the issue that the NRC (2014) committee highlighted as a research priority, namely, identifying methods of inducing conservative responding without reducing discriminability (i.e., without making it harder for eyewitnesses to tell the difference between innocent and guilty suspects).

Under some scenarios, an effect of instructions on discriminability would be apparent even in the absence of ROC analysis. For example, if the use of conservative instructions happened to selectively reduce the false ID rate while having no effect on—or even increasing—the correct ID rate, no further analysis would be needed to determine the effect of those instructions on discriminability. Instead, outcomes like that would unambiguously indicate that conservative instructions *increase* discriminability. In a meta-analysis of this literature, Steblay (1997) concluded that, compared with biased instructions, unbiased instructions selectively reduce the false ID rate while having no effect on the correct ID rate. However, Clark's (2005) re-analysis of the same data found that both the correct ID rate and the false ID rate decreased following unbiased instructions, as would be expected if different instructions induce different levels of response bias. Such findings leave unanswered the question of how biased versus unbiased instructions affect discriminability, if at all.

Confidence-based and instruction-based ROCs should be similar

There is no a priori reason to think that confidence-based and instruction-based ROCs will differ from each other. For example, the most conservative point on the ROC could be obtained either by counting suspect IDs only if they were made with 100% confidence (the usual approach) or by instructing participants not to make an ID unless they are 100% confident that the identified individual is the guilty

suspect. Logically, these two strategies should yield the same ROC point, and the same should hold true for any level of expressed confidence versus instructed confidence.

Of course, empirically, the two methods for generating ROC data might not yield exactly the same ROC points. However, even in that case, it seems reasonable to suppose that discriminability would be the same in either case. In other words, the points generated by the two methods would be expected to fall along the same ROC curve even if those points did not fall directly atop one another as they logically should. Manipulations that *would* be expected to affect discriminability are those that affect the strength of the memory trace, such as exposure duration, retention interval, lighting, and so on. In contrast to manipulations like that, there is no a priori reason to expect that different ways of varying response bias would also affect discriminability. Indeed, in the basic list memory recognition literature, and also in the basic perception literature, the confidence-based and instruction-based data have often been found to trace out essentially the same ROC curve (e.g., Dubé & Rotello, 2012; Egan, Schulman, & Greenberg, 1959; Koen & Yonelinas, 2011).

These findings are largely consistent with other findings from the basic memory and perception literatures in which ROCs constructed in a variety of ways (beyond comparing instructions vs. confidence ratings) are usually similar even though small differences are sometimes observed (e.g., Benjamin, Tullis, & Lee, 2013; Swets, Tanner, & Birdsall, 1961). For example, Swets et al. used a visual perception task in which a circular stimulus was briefly presented (or not). ROCs were constructed using either confidence ratings or by manipulating response payoffs in different conditions (analogous to using instructions to manipulate response bias in different conditions). They reported that the confidence-based ROC was slightly but consistently lower than was the payoff-based ROC in four participants they tested. However, they concluded that the small differences they observed were more likely due to methodological issues (e.g., all participants were tested in the payoff condition first, followed by the confidence condition) than to any real difference between the two ROC methods. Benjamin et al. did not compare different ROC methods but found that the more response options there were on the confidence scale, the lower the observed ROC. Such a finding might explain why the confidence-based ROC sometimes falls below an ROC generated using other methods. Again, however, the effects they observed were small.

The key point is that there is no logical or empirical reason to believe that the two ROC methods will yield substantially different results when an eyewitness ID paradigm is used. Nevertheless, as a general rule, policymakers are unwilling to make the leap of faith and presume that results from list memory studies automatically apply to eyewitness ID. Thus, for research on instructional biasing to have any influence on policy, the issue would first have to be investigated using an eyewitness ID paradigm, as recommended by the NRC (2014) report and as we do here.

In summary, the goal of the research reported here was to empirically answer the following question: In an eyewitness ID paradigm, is the instruction-based ROC the same as the confidence-based ROC, or do different instructions yield points that fall above or below the confidence-based ROC curve? Currently, there is no information available to answer that question. As noted earlier, it is an important applied question because, for example, a jurisdiction that is interested in inducing more conservative responding in eyewitnesses has at least two options: either count only suspect IDs made with relatively high confidence or induce a more conservative response bias using an instructional manipulation. Are these two options as interchangeable as they should be, or is one better than the other?

Table 1. Number of participants tested in each condition (total n = 5141 after excluding 82 participants who answered the validation question incorrectly)

Condition	Target absent	Target present	Total
Confidence	488	490	978
Liberal	537	529	1066
Neutral	516	521	1037
Unbiased	486	498	984
Conservative	539	537	1076

METHOD

Participants

Participants (n = 5223) were recruited from universities across the USA and Amazon Mechanical Turk (MTurk n = 736; www.mturk.com). There were 3587 female participants (69%), 1613 male participants (31%), and 23 (<1%) participants who did not indicate gender (age, M = 22.20 years, SD = 7.21, *range* 18–70 years). University students received course credit for their participation, and MTurk workers received $0.20 for their participation.

Participants were randomly assigned to either a confidence rating condition (n = 995) or one of four instructional biasing conditions (n = 4228; liberal, neutral, unbiased, or conservative, defined later), and a target-present (n = 2622) or target-absent (n = 2601) lineup. Eighty-two of the participants were excluded from analysis because they did not correctly answer the validation question (described in the Procedure section). Table 1 presents the number of participants assigned to each condition who were included in the final analysis.

Materials

The study stimulus (a brief video of a mock crime) and test stimuli (photos of the culprit and matched fillers) were the same as those used in Mickes, Flowe, and Wixted (2012). The video showed the culprit, a 22-year-old White man, walking past an unoccupied office and stealing a laptop.[1] Six-person simultaneous lineups (displayed in a 2 × 3 array) contained the culprit (target present) or did not contain the

[1] Using only one video may cause some concern that the results are less generalizable to the real world than they otherwise would be because, in the real world, different witnesses see different perpetrators. We have, however, analyzed ROC data from many studies using different procedures that used a single perpetrator and procedures that used multiple perpetrators from different labs and have never seen any notable difference in terms of the issues addressed here.

culprit (target absent). The five fillers in target-present lineups and the six fillers in target-absent lineups were White men who matched the description of the culprit. Following Mickes et al. (2012), for each participant, the fillers were culled from the Corrections Offender Network database (www.dc.state.fl.us). The culprit and fillers were randomly positioned per lineup.

The descriptive names of the four biasing conditions (liberal, neutral, unbiased, and conservative) are worth clarifying because they differ slightly from prior usage. For example, Wells, Smalarz, and Smith (2015a) stated that

> Biased lineup instructions are those that either fail to warn the witness that the culprit might not be in the lineup or imply that the culprit is in the lineup. Unbiased instructions, in contrast, warn the witness that the culprit might not be in the lineup (p. 109).

In our study, we had one condition that failed to warn the witness that the culprit might not be in the lineup and another that implied that the culprit was in the lineup. Although both of these instructions correspond to the Wells et al. definition of 'biased' instructions, we separately labeled them as 'neutral response bias instructions' (no warning that the culprit may or may not be in the lineup) and 'liberal response bias instructions' (which implied that the culprit was in the lineup). The term 'liberal response bias instructions' was used in preference to the more common label 'biased instructions', which does not indicate the direction in which responding is biased.

In addition to those two instructional conditions, we also included two other instructional conditions. One used 'unbiased instructions', which indicated that the culprit may or may not be in the lineup. The other used 'conservative response bias instructions', which implied, if anything, that the culprit is not in the lineup and that any ID should be made only if certainty is high. The key phrases in the instructions for each of the five conditions (including the confidence condition) were as follows:

*Confidence instruction*s: 'The person from the video may or may not be in the lineup. If you see the person from the video, please pick him; otherwise, choose the "not present" option'. Thus, unbiased instructions were used in this condition.

Liberal response bias instructions: 'Too many witnesses choose the "not present" option even when the person who committed the crime is in the lineup. It would be better to pick someone instead, even if you are not sure. Please choose the person you think is most likely to have appeared in the video unless you are 100% certain the person you saw is not in the lineup.'

Neutral response bias instructions: 'If you see the person from the video in the lineup, please pick him; otherwise, choose the "not present" option.'

Unbiased instructions: 'The person from the video may or may not be in the lineup. If you see the person from the video, please pick him; otherwise, choose the "not present" option.'

Conservative response bias instructions: 'Too many innocent people have been wrongly convicted because they were incorrectly chosen from a lineup. It would be better to choose the "not present" option than to pick someone when you are not certain of your choice. Please choose the "not present" option unless you are 100% certain the person you saw is in the lineup.'

Procedure

Participants were instructed to pay special attention to the video because they would have to answer questions about it later. The video was followed by a 5-minute distractor task (a game of Tetris). Next, participants were presented with instructions (that were pre-recorded and played while they were also displayed on the screen) based on their condition. They were then presented with a lineup (either target present or target absent) and made their decision (i.e., they either chose someone from the lineup or chose the 'not present' option). For those in the confidence condition, after an ID was made, they provided a rating of their confidence (where 0 = *guessing* and 100% = *absolutely certain*). All participants were then asked filler questions about the video and the validation question, 'What crime did the perpetrator commit?' All questions were four-option, multiple-choice questions.

RESULTS

The α level for all statistical tests was .05. Figure 1 presents the basic findings from the confidence condition and, separately, from the four biasing conditions. The confidence-based ROC data are shown as filled gray circles. The right-

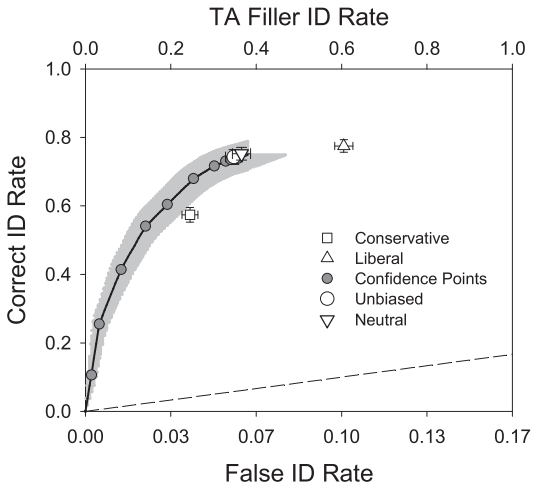

Figure 1. Receiver operating characteristic data from the confidence condition (filled circles) as fit by PROC software (solid black function, with estimated standard errors of the fit shown in gray). The four open symbols represent the correct and false ID rates from the four biasing conditions (upright triangle = liberal instructions, inverted triangle = neutral instructions, circle = unbiased instructions, and square = conservative instructions). Error bars on the open symbols represent standard errors. The dashed diagonal line represents chance performance. Note that the top horizontal axis shows the overall filler ID rate from target-absent lineups (i.e., filler IDs counted from target-absent lineups divided by the total number of target-absent lineups), whereas the ;bottom horizontal axis shows the estimated false (suspect) ID rate by dividing that value by the lineup size of 6, which is the typical strategy for estimating the false ID rate when a fair target-absent lineup is used.

most confidence symbol represents the correct and false ID rate obtained when all suspect IDs are counted regardless of confidence. Thus, this point is what is typically reported as the overall correct and false ID rates, and it represents the data from which a diagnosticity ratio is typically computed. Each additional correct and false ID rate down and to the left (i.e., each additional point on the confidence-based ROC) was computed after setting an ever-higher standard on the confidence scale for counting IDs. The second point to the left, for example, was computed by not counting any suspect IDs from target-present lineups or any filler IDs from target-absent lineups made with a confidence rating of 10 or less on the 100-point confidence scale.

Note that the top horizontal axis in Figure 1 shows the overall filler ID rate from target-absent lineups (i.e., filler IDs counted from target-absent lineups divided by the total number of target-absent lineups), whereas the bottom horizontal axis shows the estimated false (suspect) ID rate by dividing that value by the lineup size of 6, which is the typical strategy for estimating the false ID rate when a fair target-absent lineup does not have a designated innocent suspect (e.g., Clark, Moreland, & Gronlund, 2014). Obviously, the ROC data are not affected by the decision to use one incorrect ID rate or the other. The solid curve connecting the confidence data (i.e., connecting the filled gray circles) represents the ROC curve as estimated by the PROC software package (Robin et al., 2011). This is an atheoretical curve that basically connects the data points (i.e., the curve is not estimated based on any theoretical assumption about recognition memory), usually for the purpose of computing the partial area under the ROC. Here, however, our concern is with the general trajectory of the ROC (i.e., with the ROC curve itself), not the area under it. The estimated vertical and horizontal standard errors of the ROC curve (based on 10 000 bootstrap trials) are shown in gray.

The four open symbols represent the correct and false ID rates from the four different instructional biasing conditions. The data from the neutral and unbiased conditions unexpectedly turned out to be nearly identical, but the data from the liberal and conservative biasing conditions differed from the other conditions in the expected directions. In the liberal condition, the correct and false ID rates were both high relative to the neutral/unbiased instruction conditions. In the conservative condition, the correct and false ID rates were both low relative to the neutral/unbiased instruction conditions. The ROC points from the four biasing conditions, which constitute the instruction-based ROC, appear to fall on, or at least near, the confidence-based ROC. The ROC point from the conservative condition is a possible exception in that it falls somewhat below the confidence-based ROC data. The liberal ROC point might be an exception as well, but it is hard to tell without theoretically extrapolating the confidence-based ROC. Nevertheless, for the most part, the ROC path traced out by the four instructional biasing conditions appears to be similar to the path traced out by the confidence ratings.

In Figure 1, and as noted earlier, the conservative point from the instructional biasing condition falls below the confidence-based ROC to such an extent that the standard errors from the two conditions do not overlap. We know of no test that would indicate whether or not that effect is statistically significant, but we can safely conclude that there is at least a trend in that direction and, more conclusively, that using conservative biasing instructions does not *increase* discriminability compared with the confidence-based ROC. Instead, if anything, the confidence-based ROC yields higher discriminability in the more conservative region of the ROC.

We next separately analyze the confidence-based and instruction-based ROC data to determine whether or not they exhibit similar trends with respect to the diagnosticity ratio by computing its value for each point on the ROC. In the field of medicine, the diagnosticity ratio is called the *positive* likelihood ratio. In accordance with standard medical terminology, we will henceforth refer to the diagnosticity ratio for eyewitness ID as the positive diagnosticity ratio (DR+). The positive diagnosticity ratio is equal to correct ID rate/false ID rate. To illustrate why DR+ is not a useful measure of overall diagnostic accuracy, we also compute the *negative* diagnosticity ratio (DR−) for each point on the ROC. The negative diagnosticity ratio is equal to (1−correct ID rate)/(1−false ID rate). Like the positive diagnosticity ratio, the negative diagnosticity ratio (multiplied by the prior odds of a target-present lineup) indicates the odds that a suspect is guilty, except that now the measure applies to those who are *not* identified. Thus, a higher negative diagnosticity ratio reflects a less trustworthy non-ID. As with the positive diagnosticity ratio, the more conservative the decision criterion becomes, the higher the value of the negative diagnosticity ratio becomes. Computing both the DR+ and the DR− for each point on the ROC illustrates the inherent trade-off associated with manipulating response bias (thereby illustrating why a higher DR+ per se is not an indication of overall diagnostic superiority).

Confidence-rating ROC analyses

Table 2 shows, for each level of confidence, the frequency counts of suspect and filler IDs from target-present lineups and filler IDs from target-absent lineups. The data from

Table 2. Frequency counts of suspect IDs (SIDs), filler IDs (FIDs), and lineup rejections (no IDs) made from target-present and target-absent lineups in the confidence-based receiver operating characteristic condition

	Target present			Target absent	
	SIDs	FIDs	No IDs	FIDs	No IDs
Confidence					
0–30	11	4		26	
40	18	4		24	
50	37	4		30	
60	31	11		25	
70	62	6		28	
80	78	4		25	
90	73	2		9	
100	52	0		7	
ID sum	362	35	93	174	314
Total		490		488	

Note: Confidence for 0–30 was collapsed because there were few responses in that confidence range.

confidence levels of 0 through 30 were collapsed because there were very few suspect or filler IDs from target-present lineups in that range. Table 3 shows performance measures computed from the frequency data shown in Table 2. The correct ID rate for a given level of confidence is equal to the number of suspect IDs from target-present lineups made with that level or a higher level of confidence divided by the total number of target-present lineups, N_{TP}. The false ID rate for a given level of confidence is the number of filler IDs from target-absent lineups made with that level or a higher level of confidence divided by the total number of target-absent lineups, N_{TA}, and then divided again by lineup size (to estimate the false suspect ID rate).

As expected, the correct ID rate decreases as responding becomes more conservative (i.e., as a higher criterion level of confidence is used to count IDs), and so does the false ID rate (Table 3). In addition, DR+ increases as responding becomes more conservative, but DR− increases as well. The increasing positive diagnosticity ratio means that the odds that an identified suspect is guilty increase as responding becomes more conservative. Similarly, the increasing negative diagnosticity ratio means that the odds that a non-identified suspect is guilty also increase as responding becomes more conservative. Thus, these values depict the trade-off associated with more liberal versus more conservative responding (Clark, 2012). These findings with respect to DR+ replicate trends observed in recent studies that reported confidence-based ROC data (e.g., Gronlund et al., 2012; Mickes et al., 2012).

Instruction-based ROC analyses

Table 4 presents the instructional biasing data. The table shows the number of suspect IDs, filler IDs, and lineup rejections (i.e., no IDs) from target-present and target-absent lineups in each of the four biasing conditions. The data are arranged from liberal to conservative responding (top row to bottom row). Note that, unsurprisingly, as responding becomes increasingly conservative, the number of suspect IDs and filler IDs from target-present lineups decreases, whereas the number of no IDs increases. A 2 × 4 chi-square test of IDs (suspect IDs + filler IDs) versus no IDs from target-present lineups across the four biasing conditions was highly significant, $\chi^2(3) = 155.0$, $p < .001$. Similarly, for target-absent lineups, the number of IDs (filler IDs only, because there was no designated innocent suspect) decreased and the number of no IDs increased as responding becomes more conservative, an effect that was also highly significant, $\chi^2(3) = 156.7$, $p < .001$. These results indicate that the instructional biasing manipulation had the expected effect on response bias.

Table 3. Performance measures computed from the frequency counts shown in Table 2

Confidence	Correct ID rate	False ID rate	DR+	DR−
0–30	0.74	0.06	12.4	0.28
40	0.72	0.05	14.2	0.30
50	0.68	0.04	16.0	0.33
60	0.60	0.03	18.8	0.41
70	0.54	0.02	22.9	0.47
80	0.41	0.01	29.6	0.59
90	0.26	0.01	46.7	0.75
100	0.11	0.002	44.4	0.90

Note: The correct and false ID rates are cumulative in that all IDs made with the indicated level of confidence *or higher* (e.g., 40 or higher for the second row of data) were counted as IDs.
DR+ = positive diagnosticity ratio; DR− = negative diagnosticity ratio.

Table 4. Frequency counts of suspect IDs (SIDs), filler IDs (FIDs), and lineup rejections (no IDs) made from target-present and target-absent lineups in the instruction-based receiver operating characteristic condition

	Target present				Target absent		
Condition	N_{TP}	SIDs	FIDs	No IDs	N_{TA}	FIDs	No IDs
Liberal	529	410	71	48	537	325	212
Neutral	521	392	45	84	516	189	327
Unbiased	498	370	34	94	486	169	317
Conservative	537	308	22	207	539	132	407

Note: For each biasing condition, N_{TP} is the total number of target-present lineups, and N_{TA} is the total number of target-absent lineups.

Table 5 presents performance measures associated with the four instructional biasing conditions. These performance measures were computed from the observed data shown in Table 4. The correct ID rate is equal to the number of suspect IDs from target-present lineups divided by the total number of target-present lineups, N_{TP}, whereas the false ID rate is the number of filler IDs from target-absent lineups divided by the total number of target-absent lineups, N_{TA}, and then divided again by lineup size (to estimate the false suspect ID rate). As expected, the correct ID rate decreases as responding becomes more conservative, and so does the false ID rate. Importantly, as shown in Table 5, DR+ and DR− both increase as responding becomes more conservative. Thus, the trends that are observed in the confidence-based ROC data are also observed in the instruction-based ROC data.

Signal detection model fit

As noted earlier, the liberal instructional biasing ROC point (like the conservative point) may fall below the confidence-based ROC, but there is no way to extrapolate the confidence-based ROC without the use of a theory. We therefore fit the confidence-based ROC data shown in Table 2 with what is arguably the simplest signal detection model that can be applied to lineups. This model was described by Wixted and Mickes (2014) and was recently fit to ROC data by Colloff et al. (2016). According to this model, memory strength values for lures (innocent suspects and fillers for a fair lineup) and for targets (guilty suspects) are distributed according to Gaussian distributions with means of μ_{Lure} and μ_{Target}, respectively, and standard deviations of σ_{Lure} and σ_{Target}, respectively. A six-member target-present lineup is conceptualized as five random draws from the lure distribution and one random draw from the target distribution, and a fair six-member target-absent lineup is conceptualized as six random draws from the lure distribution. Using the simplest decision rule, we made an ID to the individual in a lineup with the greatest memory

Table 5. Performance measures across the four biasing conditions computed from Table 4

Condition	Correct ID rate	False ID rate	DR+	DR−
Liberal	0.78	0.10	7.68	0.25
Neutral	0.75	0.06	12.33	0.26
Unbiased	0.74	0.06	12.82	0.27
Conservative	0.57	0.04	14.05	0.44

Note: Unlike the confidence data in Table 3, these correct and false ID rates are not cumulative.
DR+ = positive diagnosticity ratio; DR− = negative diagnosticity ratio.

strength, assuming that strength at least exceeds the decision criterion for making an ID with the lowest level of confidence. According to this model, each level of confidence is associated with its own decision criterion, and the confidence associated with an ID corresponds to the highest confidence criterion exceeded by the memory strength of the most familiar face in the lineup (whether that face is a suspect or a filler). When fit to data produced by many participants, each of whom provided a single confidence rating, the model conceptualizes group performance (not the performance of any single participant).

By convention, μ_{Lure} is set to 0 and $\sigma_{\text{Lure}} = 1$. The remaining parameters—μ_{Target}, σ_{Target}, and eight confidence criteria (one for each point on the confidence-based ROC)—were estimated by adjusting them until the chi-square comparing observed and predicted values was minimized (using fminsearch in MATLAB). The fit was reasonably good, although not perfect, $\chi^2(14) = 28.58$, $p = .012$ (i.e., the deviations from the best-fitting model are significant). The smooth curve in Figure 2A shows the predicted ROC curve from the best-fitting model, and it is clear that the model captures the basic trends in the data. Thus, while a more complex model may be needed to fit the data exceptionally well, a simple signal detection model appears to be a useful tool for conceptualizing the basic trends in ROC data generated from a lineup. Figure 2B again shows the predicted ROC curve for the confidence-based ROC data, but this time it is drawn through the four instruction-based ROC points. This plot reinforces the interpretation of the data presented earlier: the instruction-based ROC data are similar to the confidence-based ROC data, but the conservative instruction-based ROC point falls below the confidence-based ROC curve, and, apparently, so does the liberal instruction-based ROC point (although to a lesser extent).

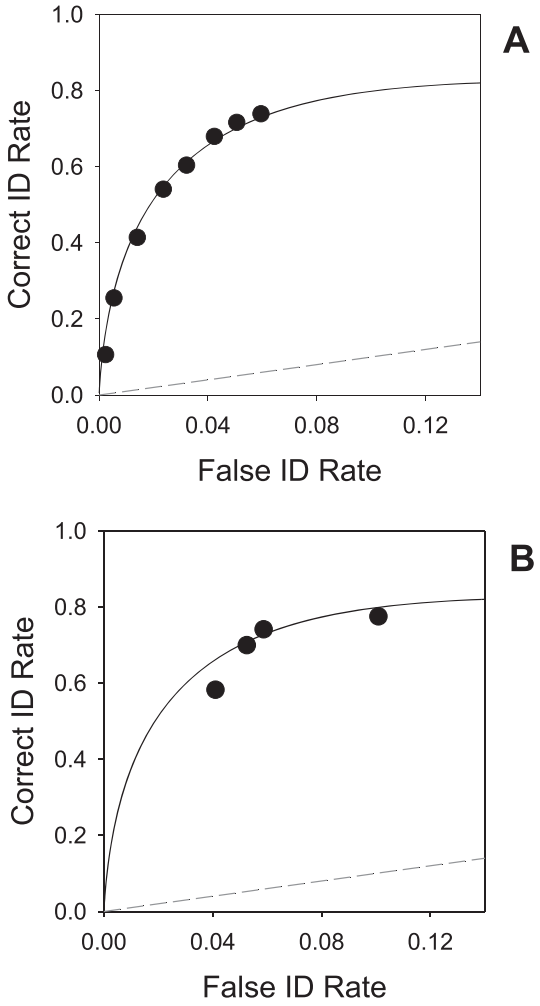

Figure 2. (A) Fit of the simple signal detection model to the confidence-based receiver operating characteristic (ROC) data (the smooth curve shows the predicted values from the best-fitting model). (B) Instruction-based ROC data from the four biasing conditions. The smooth curve is the same curve fit to (and drawn through) the confidence-based ROC data in panel A.

DISCUSSION

In this investigation of simultaneous lineup performance, we compared ROC data obtained from two traditional methods from the basic recognition and perception literatures: (i) confidence ratings and (ii) instructions designed to induce different levels of response bias. As shown in Figure 1, the results suggest that the family of correct and false ID rates computed from confidence ratings (i.e., the confidence-based ROC), and the family of correct and false ID rates generated by manipulating response bias across conditions (i.e., the instruction-based ROC), fall on approximately the same curve. However, the ROC point from the conservative instructional biasing condition deviates from the confidence-based ROC in two notable ways: first, it falls slightly below the confidence-based ROC curve (as the liberal ROC point appears to do as well), and, second, it falls at a much more liberal position on the ROC than it logically should. As noted earlier, participants in that condition were instructed to make an ID only if they were 100% confident that the identified individual was the guilty suspect. Even so, that conservative ROC point falls well to the right of the corresponding (leftmost) point on the confidence-based ROC (i.e., it falls in the vicinity of the point on the confidence-based ROC that corresponds to approximately 50% to 60% confidence). According to the fit of a signal detection model, the liberal ROC point also appears to fall below the confidence-based ROC curve. We next consider the possible theoretical implications of these effects and then consider possible policy implications of our findings.

Theoretical considerations

Criterion/instructional variability

One possible explanation for the seemingly anomalous conservative ROC point from the instructional biasing condition is that it reflects a phenomenon analogous to *criterion variability*. As recently noted by Benjamin, Diaz, and Wee (2009), if an individual participant's decision criterion varies from trial to trial on a list memory recognition test involving many test trials (i.e., if a somewhat liberal response bias were in effect on some trials, but a somewhat conservative response bias were in effect on other trials), the result would show up as reduced discriminability in the form of a lower ROC than would otherwise be observed. A similar explanation may apply to ROC data in which participants are each tested on only one trial.

Conceivably, participants may differ in the degree to which they comply with instructions to refrain from making an ID unless they are sufficiently confident. If so, that additional source of variance would have the effect of reducing discriminability for the same reason that criterion variability within a single participant tested across many recognition trials reduced discriminability (Benjamin et al., 2009). For example, if the instructions in the conservative condition cause some participants not to make an ID unless they are at least 100% confident (in accordance with the instructions), others not to make an ID unless they are at least 55% confident, and still others not to make an ID unless they are at least 10% confident, then the correct and false ID rates in the conservative condition would decrease relative to the more neutral biasing conditions, but they would not decrease as much as they should. In other words, the conservative ROC point would not shift to the left all the way to the leftmost point on the confidence-based ROC (which corresponds to 100% confidence). Moreover, owing to that variability in cooperating with the instructions, the correct and false ID rates in the conservative condition would also now fall on a lower ROC compared with the confidence-based ROC. Similar considerations would help to explain why the correct and false ID rates in the liberal condition also appear to fall on a lower ROC compared with the confidence-based ROC (Figure 2B).

Other explanations for why the conservative ROC points fell on a lower ROC are certainly possible. For example, extreme biasing instructions might cause participants to pay less attention to the task at hand as they devote attentional resources trying to comply with the instructions. The explanation offered earlier has the advantage of being conceptually related to prior accounts of why confidence-based and instructional-biasing ROCs sometimes differ, but further research would be needed to establish its validity over other possible interpretations.

Underlying (theoretical) discriminability versus empirical discriminability

In recent years, a lively debate has emerged over the use of ROC analysis for testing lineup performance (Lampinen, 2016; Rotello & Chen, 2016; Smith, Wells, Lindsay, & Penrod, 2017; Wells, Smalarz, & Smith, 2015a; Wells, Smith & Smalarz, 2015b; Wixted & Mickes, 2015a, 2015b; Wixted, Mickes, Wetmore, Gronlund, & Neuschatz, in press). However, the debate is largely focused on the theoretical issue of *underlying* discriminability. For example, in their most recent critique of ROC analysis, Smith et al. (2017) focus exclusively on the issue of theoretical discriminability. It is therefore important to emphasize that the policy-related implications of the work reported here are not in any way related to the issue of underlying theoretical discriminability (e.g., discriminability as assessed by fitting a signal detection model). Instead, the policy-related implications derive solely from our assessment of *empirical* discriminability and *empirical* trends that are observed when ROC data are collected.

What is the difference between theoretical and empirical discriminability? We noted earlier that empirical ROC data that bow further up and away from the diagonal line of chance performance indicate higher discriminability in the sense that eyewitnesses are better able to tell the difference between innocent and guilty suspects. That interpretation applies to empirical reality, not to any theoretical interpretation (e.g., it does not rely on any interpretation provided by any signal detection model that might be fit to the data). Empirically, the more the ROC data pull above the diagonal line of chance performance, the better able eyewitnesses are to objectively sort innocent and guilty suspects into their correct categories. Only that empirical reality has any applied implications with respect to correct and incorrect suspect IDs because policymakers are concerned with what is actually (i.e., empirically) achievable, not with what theoretically might be the case.

Although empirical and theoretical discriminability typically go hand in hand, they do not necessarily have to agree about which of two conditions yields higher empirical discriminability. Two conditions can differ in empirical discriminability yet not differ at all in terms of underlying discriminability (i.e., in terms of what a particular model assumes). For example, an objective empirical ROC advantage has been reported for simultaneous lineups over showups (Wetmore, Neuschatz, Gronlund, Wooten, Goodsell & Carlson, 2015b), a finding that has clear policy implications. By contrast, the same results can be mimicked by a model that assumes that underlying theoretical discriminability is the same for the two procedures (e.g., Lampinen, 2016; Smith et al., 2017). However, that fact has no policy implications whatsoever. Which of two competing theoretical models is more viable (one that assumes that theoretical discriminability differs in the same way that empirical discriminability does vs. one that assumes a dissociation between theoretical and empirical discriminability) is a matter for theoreticians to debate. The same consideration applies to the signal detection analysis summarized here in Figure 2A and B. According to the specific model we fit to the confidence-based ROC data, the liberal ROC point from the instructional biasing condition falls somewhat below the confidence-based ROC. However, a different signal detection model might lead to a different conclusion. Again, that is a matter for theoreticians to debate, not for policymakers to worry about. With regard to the conservative ROC point from the instructional biasing condition, no theory is needed to see that (and no

theory can change the fact that), if anything, it falls below the confidence-based ROC.

The present results address the recent ROC controversy from another angle as well. Although not conceptualized as such, prior work on biased versus unbiased instructions is itself an example of ROC analysis, with two points being generated on the instruction-based ROC (one point from the biased condition and a second more conservative point from the unbiased condition). That fact is worth mentioning because in the ongoing controversy over the validity of ROC analysis for lineups, some of the main opponents include those who have conducted ROC analysis in previous research simply by computing correct and false ID rates across several conditions that used instructions to manipulate response bias (e.g., Cutler, Penrod, & Martens, 1987; Lindsay et al., 1991). The main concern that has been raised about ROC analysis is that it focuses on suspect IDs without regard for filler IDs. However, prior work on biased versus unbiased instructions also computed correct and false suspect ID rates, without regard for filler IDs. Computing correct and false suspect ID rates across different levels of response bias, which is what has been carried out in prior work on biased versus unbiased instructions, *is* ROC analysis. Thus, to the extent that ROC analysis is judged to be inappropriate for lineups, the same judgment would have to apply to prior research comparing the effects of different lineup instructions.

In our view, there is nothing inappropriate about measuring correct and false suspect ID rates for different levels of response bias using either instructions or confidence ratings. In fact, no pair of correct and false ID rates obtained from a single lineup procedure is sacrosanct. If it is legitimate to compute one pair of correct and false ID rates (as nearly every study of lineup performance ever conducted has performed), then it is equally legitimate to compute all of the correct and false ID rate pairs that make up the ROC because they all have equal standing. This is true whether the ROC points are generated using confidence ratings or instructions. Conceivably, the instruction-based ROC data we have reported here will communicate that critical point more clearly than past work on confidence-based ROC data has.

Potential policy implications

Prior research has generally been interpreted to mean that unbiased instructions (i.e., instructions that are neutral with respect to presence or absence of the perpetrator in the lineup) are objectively superior to biased instructions (i.e., instructions that imply that the perpetrator is in the lineup and therefore induce liberal responding). However, the measure that has often been used to make that determination is the diagnosticity ratio (Wells & Lindsay, 1980), which we have represented here as DR+. This logic always favors more conservative responding over less conservative responding because more conservative responding will always yield a higher positive diagnosticity ratio (cf. Rotello, Heit, & Dubé, 2015). For example, using this logic, the conservative instructional biasing condition in our study—which yielded the highest positive diagnosticity ratio—should be preferred to the other instructional biasing conditions (Table 5), including the unbiased condition.

In truth, it would be a mistake to claim that the conservative instructional biasing condition is objectively superior to the other instructional biasing conditions because that assessment involves a subjective value judgment about the optimal balance between DR+ and DR− (cf. Clark, 2012). Most would probably agree (ourselves included) that responding on an eyewitness ID procedure ought to be conservative in the objective sense that the false suspect ID rate should be low even if it means that the miss rate (equal to 1−correct suspect ID rate) is not commensurately low. Indeed, in Table 5, it is clear that responding is more conservative than liberal in that objective sense for all four response bias conditions (including the 'liberal' condition). That is, in each condition, 1−correct ID rate is greater than the corresponding false suspect ID rate. However, no matter which condition you start with, inducing more conservative responding will always yield a higher DR+. This was already known to be true for confidence-based ROC data, and our findings suggest that the same appears to be true for instruction-based ROC data.

In terms of policy implications, what the data suggest is that if, based on cost/benefit analysis, a jurisdiction wanted to induce more conservative responding than is achieved by the use of neutral/unbiased instructions alone (e.g., in a jurisdiction willing to convict solely on the basis of eyewitness evidence), it might be better to use neutral instructions in conjunction with confidence ratings to achieve that outcome. For example, using neutral instructions in conjunction with the 50% confidence criterion (sixth point from the left on the confidence-based ROC) would yield a false ID rate comparable with that associated with the conservative instructional biasing condition while achieving a noticeably higher correct ID rate (i.e., better discriminability). Another advantage of using this approach is that would allow a jurisdiction to use multiple levels of response bias, perhaps a relatively liberal one (e.g., ≥50%) for purposes of deciding whether or not to further investigate a suspect, and a much more conservative one (e.g., ≥90%) for purposes of deciding whether or not to bring charges against a suspect. This approach might be a viable alternative to a police strategy that treats any suspect ID as a 'positive ID' (without any consideration given to confidence).

Finally, it is worth noting that, in one important respect, our findings agree with past research suggesting that biased instructions (the liberal condition here) may be inferior to unbiased instructions. In Figure 2B, there is at least a hint that the liberal condition results in reduced discriminability compared with the neutral/unbiased conditions. Thus, nothing we report here should be taken as a reason to dispute the longstanding recommendation in favor of unbiased instructions over biased (liberal) instructions. Prior research on that issue relied on an inappropriate dependent measure, but it did not necessarily reach the wrong conclusion. However, lineup instructions used by some police departments often go well beyond unbiased/neutral instructions and, if anything, tilt even more in the conservative direction. For example, two recent police department field studies made use of instructions that are often advocated by eyewitness ID

researchers (G. Wells, Steblay & Dysart, 2011, 2015c; W. Wells, 2014). These instructions emphasize the following four points to eyewitnesses who are about to view a lineup:

1. The person who committed the crime may or may not be in the lineup.
2. The investigation will continue whether or not someone is identified from the lineup.
3. It is just as important to remove suspicion from an innocent suspect as it is to convict a guilty suspect.
4. An identification does not have to be made from the lineup.

All of these instructions, if they had any effect at all, would tend to induce a more conservative response bias compared with the response bias that would otherwise be in effect (e.g., if unbiased or neutral instructions were used). Whether or not these particular instructions have a negative (or, perhaps, a positive) effect on discriminability is not known. Investigating that issue would seem to be a reasonable next step, one that would be in accordance with one of the research priorities specified in the NRC (2014) report.

ACKNOWLEDGMENTS

This work was supported in part by the Economic and Social Research Council (ES/L012642/1) to Laura Mickes and John T. Wixted and by the National Science Foundation (SES-1456571) to John T. Wixted. The content is solely the responsibility of the authors and does not necessarily reflect the views of the Economic and Social Research Council or of the National Science Foundation.

REFERENCES

Anderson, S., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36–40.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*, 84–115.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601–1608.

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *3*, 45–53.

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

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

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review*, *21*, 251–267.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, *27*, 1227–1239.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identification: The role of system and estimator variables. *Law and Human Behavior*, *11*, 233–258.

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

Dubé, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 130–151.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America*, *31*, 768–773.

Flowe, H. D., Klatt, T., & Coloff, M. F. (2014). Selecting fillers on emotional appearance improves lineup identification accuracy. *Law and Human Behavior*, *38*, 509–519.

Flowe, H. D., Smith, H. M. J., Karoğlu, N., Onwuegbusi, T., & Rai, L. (2015a). Configural and component processing in simultaneous and sequential lineup procedures. *Memory*, *24*, 306–314.

Flowe, H. D., Smith, H. M. J., Karoğlu, N., Onwuegbusi, T., & Rai, L. (2015b). Configural and component processing in simultaneous and sequential lineup procedures. *Memory*, *24*, 334–347.

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

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

Humphries, J. E., & Flowe, H. D. (2015). Receiver operating characteristic analysis of age-related changes in lineup performance. *Journal of Experimental Child Psychology*, *132*, 189–204.

Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J. L., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. *Psychology, Crime and Law*, *21*, 871–889.

Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: Bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory*, *18*, 519–522.

Lampinen, J. M., Erickson, W. B., Moore, K. N., & Hittson, A. (2014). Effects of distance on face recognition: Implications for eyewitness identification. *Psychonomics Bulletin & Review*, *21*, 1489–1494.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition*, *5*, 21–33.

Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology*, *76*, 796–802.

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

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness. *Journal of Applied Research in Memory and Cognition*, *18*, 361–376.

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

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

Neuschatz, J. S., Wetmore, S. A., Key, K., Cash, D., Gronlund, S. D., & Goodsell, C. A. (2016). Comprehensive evaluation of showups. In M. K. Miller, & B. H. Bornstein (Eds.), *Advances in psychology and law*, (Vol. *1*). New York: Springer.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and Sto analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rotello, C. M., & Chen, T. (2016). ROC analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*, 944–954.

Seale-Carlisle, T. M., & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science.*, https://doi.org/10.1098/rsos.160300.

Smith, H. M. J., & Flowe, H. D. (2015). ROC analysis of the verbal overshadowing effect: Testing the effect of verbalisation on memory sensitivity. *Applied Cognitive Psychology*, *29*, 159–168.

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, *41*, 127–145.

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law and Human Behavior*, *21*, 283–297.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340.

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness non-identifications. *Psychological Bulletin*, *88*, 776–784.

Wells, G. L., Smalarz, L., & Smith, A. M. (2015a). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory & Cognition*, *4*, 313–317.

Wells, G. L., Smith, A. M., & Smalarz, L. (2015b). 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.

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

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015c). Double-blind photolineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior*, *39*, 1–14.

Wells, W. (2014). The Houston Police Department eyewitness identification experiment: Analysis and results. Retrieved from: http://www.lemitonline.org/research/projects.html

Wetmore, S., Neuschatz, J. S., Gronlund, S. D., Key, K. N., & Goodsell, C. A. (2015a). Do the clothes make the criminal? The influence of clothing match on identification accuracy in showups. *Journal of Applied Research in Memory and Cognition*, *4*, 36–42.

Wetmore, S., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015b). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*, 8–14.

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

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

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.

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.

Wixted, J. T., Mickes, L., Wetmore, S. A., Gronlund, S. D., & Neuschatz, J. S. (in press). ROC analysis in theory and practice. *Journal of Applied Research in Memory and Cognition*.