Unfair lineups make witnesses more likely to confuse innocent and guilty suspects
Colloff, Melissa; Wade, Kimberley A; Strange, Deryn

DOI: 10.1177/0956797616655789
License: None: All rights reserved

Document Version
Peer reviewed version

Citation for published version (Harvard):
Colloff, M, Wade, KA & Strange, D 2016, 'Unfair lineups make witnesses more likely to confuse innocent and guilty suspects'. Psychological Science, vol. 27, no. 9, pp. 1227-1239.
https://doi.org/10.1177/0956797616655789

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.
Unfair lineups don’t just make witnesses more willing to choose the suspect, they also make them more likely to confuse innocent and guilty suspects.

Melissa F. Colloff\textsuperscript{1}, Kimberley A. Wade\textsuperscript{1}, Deryn Strange\textsuperscript{2}

\textbf{Affiliations}

\textsuperscript{1} University of Warwick

\textsuperscript{2} John Jay College of Criminal Justice, City University of New York

Correspondence concerning this article should be addressed to Kimberley A. Wade, Department of Psychology, University of Warwick, Coventry, UK, CV4 7AL. Email: K.A.Wade@warwick.ac.uk
Abstract

Eyewitness identification studies have focused on the idea that unfair lineups, in which the suspect stands out, make witnesses more willing to identify that suspect. We asked whether unfair lineups—featuring suspects with distinctive features—also influence subjects’ ability to distinguish between innocent and guilty suspects, and their ability to judge the accuracy of their identification. In a single experiment ($N = 8925$), we compared three fair lineup techniques used by the police to unfair lineups in which we did nothing to prevent distinctive suspects from standing out. Compared to the fair lineups, doing nothing not only increased subjects’ willingness to identify the suspect, it also markedly impaired subjects’ ability to distinguish between innocent and guilty suspects. Accuracy was also reduced at every level of confidence. These results advance theory on witness identification performance and have important practical implications for how police should construct lineups when suspects have distinctive features.

**Keywords:**

eyewitness memory; lineup fairness; distinctive features; diagnostic-feature-detection
In 1986, a woman viewed a lineup and identified Leonard Callace as her attacker. She had described the attacker as a white male with reddish-blonde, afro-style hair and a full beard. But Callace—who had a full beard, and straight hair—appeared in the lineup with five men who had only moustaches. After Callace served six years in prison, DNA evidence revealed he was not the attacker. Callace’s case, and many others, highlights the importance of preventing suspects with distinctive features from standing out in lineups (see http://www.innocenceproject.org/). But why do unfair lineups impair eyewitness identification performance? Is it because unfair lineups make witnesses more willing to identify the suspect? Or is it because unfair lineups make it more difficult for witnesses to determine if the lineup contains the actual culprit? We aimed to answer these questions.

We know that suspects who stand out are prone to be selected for the wrong reasons—namely, not because they match the witness’s memory of the culprit (Wells, Rydell, & Seelau, 1993). Why? The long-standing explanation is that witnesses tend to select the person who looks most like the culprit, much like the way a student answering a multiple choice question tends to select the option that looks most like the right answer (Wells, 1984). Indeed, it is well established that when the only person who matches the witness’s description of the culprit is the suspect, the witness tends to select the suspect instead of another lineup member (Doob & Kirshenbaum, 1973; Wells, Leippe, & Ostrom, 1979). More recent reviews and meta-analyses also show that when suspects look less like the other members of a lineup, witnesses identify the suspect more often (Clark, 2012; Fitzgerald, Price, Oriet, & Charman, 2013). Two problems arise from this tendency. First, if the suspect is the culprit, the identification is correct, but not for the right reasons—much like the student who gets the correct answer but does not actually know the right answer. Second, if the suspect is not the culprit, the misidentification might send an innocent person to prison. The observation that witnesses are more willing to identify the suspect—which
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

means correctly identifying the culprit when he is present, but incorrectly identifying an innocent suspect when he is not present—can help us to understand why unfair lineups often result in misidentifications.

Yet, a new approach, the diagnostic-feature-detection model, supports an additional prediction: Unfair lineups may also impair witnesses’ ability to differentiate between the actual culprit and an innocent suspect (Wixted & Mickes, 2014). To see why, consider what happens when a witness views the members in a lineup, whether fair or unfair. The idea is that for each lineup member’s face, features combine to create a memory signal (a sense of familiarity and recollection) and the witness uses that signal to make an identification decision. Because some features differ between the culprit and an innocent suspect, they can help the witness make a better decision. For instance, Leonard Callace had straight hair, while the culprit had an afro. But other facial features are shared by the culprit and an innocent suspect, so they cannot help the witness. For instance, Callace and the culprit each had a full beard. If witnesses give weight to these shared features, their ability to distinguish between culprits and innocent suspects will suffer.

How, then, do witnesses make identifications in an unfair lineup, where only the suspect possesses the distinctive facial feature they remember—say, a full beard? To the extent witnesses do not realize the distinctive feature is unhelpful, they might erroneously weight that feature. Giving weight to an unhelpful feature will impair their ability to discriminate between real culprits and innocent suspects. Consistent with this idea, one study showed that witnesses were better able to distinguish between guilty and innocent suspects when all lineup members, including the suspect, had the same emotional expression. But witnesses found it harder to distinguish between innocent and guilty suspects when the suspect was the only one with that expression (Flowe, Klatt, & Collof, 2014). Presumably, those who saw the “matched
expression” lineup discounted the shared emotional expression and used other, useful information to make an identification. By contrast, those who saw the “unmatched expression” lineup weighted the shared emotional expression, even though it was objectively unhelpful. Other studies have found that people are better able to distinguish between innocent and guilty suspects when they are presented with a fair lineup rather than a single photo of a suspect (i.e., a showup, Key et al., 2015; Wetmore et al., 2015). Again, the fair lineup may permit subjects to discount unhelpful features but a single photo may not.

In the real-world, police guidelines for constructing lineups often state that the police should prevent suspects with distinctive features from unduly standing out. In the US, England and Wales, for instance, police sometimes artificially replicate a suspect’s distinctive feature across the lineup members (hereafter replication, see Fig. 1a); other times, they conceal the feature on the suspect and conceal a similar area on the other members (Police and Criminal Evidence Act, Code D, 1984; Technical Working Group for Eyewitness Evidence, 1999). Concealing involves either pixelating the area of the feature (hereafter pixelation, Fig. 1b), or covering the area with a solid black rectangle (hereafter block, Fig. 1c). These techniques represent a heartening translation of science into practice. Nonetheless, many efforts to make lineups fair are unsuccessful, and police officers still often do nothing, and leave these suspects to stand out (e.g. MacLin, MacLin, & Albrechtsen, 2006; Valentine & Heaton, 1999; Wogalter, Malpass, & Mcquiston, 2004).

How, then, might replication, pixelation or block lineups affect lineup performance? First, because the suspect does not unduly stand out, witnesses should be less willing to identify the suspect. Second, because the distinctive feature appears either on every lineup member (replication), or on none of the lineup members (pixelation, block), witnesses should be more likely to weight something other than the distinctive feature. Therefore, they should also be better
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

able to distinguish between the culprit and an innocent suspect. By contrast, if a suspect is left to stand out (hereafter *Do nothing* lineups, Fig. 1d), witnesses should be more willing to choose the suspect, and they should find it harder to distinguish between the culprit and an innocent suspect.

The current research tested these hypotheses.
Fig. 1.

Examples of (a) a Replication Lineup, (b) a Pixelation Lineup, (c) a Block Lineup, and (c) a Do-nothing (Unfair) Lineup. Top left image in each lineup is the suspect with the distinctive facial feature.

Method

Design

We used a 4 (lineup type: replication, pixelation, block, do-nothing) × 2 (target: present, absent) between-subjects design. Our data-collection stopping rule was to recruit as many subjects as possible before the end of spring term, with a minimum of 1000 subjects with useable data in each of the eight conditions.

Subjects

The subjects were 9841 adults from around the world who completed the task online. We excluded 916 people (10% in total; between 89-218 in each of the 8 conditions), which resulted in a total sample size of 8925. We excluded subjects who experienced technical difficulties while watching the video (n = 689, 7% in total), experienced programming errors while viewing the lineup (n = 128, 1% in total), or incorrectly answered an attention check question on the content of the video (n = 99, 1% in total). The final sample consisted of 5495 subjects recruited from social network sites who were entered into a prize draw for four £50 Amazon vouchers; 2405 subjects recruited via Amazon Mechanical Turk who received $0.60; 871 students recruited from John Jay College of Criminal Justice who received extra credit in a course; and 154 students recruited from a sixth form (final year of high school) in the UK who completed the study as part of a research methods course. Because the pattern of results was the same among the Internet and
student samples, we combined the data for our analyses. Each cell contained between 1017 and 1145 subjects. We also checked for multiple responses by the same individual by examining IP addresses and email addresses. These checks revealed 26 possible cases of duplicates (i.e. 0.003 of subjects). Our results are the same regardless of whether we include or exclude these people.

Table 1 shows a demographic breakdown of the sample.

Table 1.  
*Demographic Information For Social Media, Mechanical Turk, University and Sixth Form Samples*

<table>
<thead>
<tr>
<th></th>
<th>Social Media</th>
<th>Mechanical Turk</th>
<th>University</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1498</td>
<td>1091</td>
<td>265</td>
<td>40</td>
</tr>
<tr>
<td>Female</td>
<td>3960</td>
<td>1309</td>
<td>599</td>
<td>114</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>37</td>
<td>5</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-20</td>
<td>1606</td>
<td>79</td>
<td>593</td>
<td>149</td>
</tr>
<tr>
<td>21-30</td>
<td>1693</td>
<td>997</td>
<td>252</td>
<td>0</td>
</tr>
<tr>
<td>31-40</td>
<td>870</td>
<td>675</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>41-50</td>
<td>649</td>
<td>326</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>51-60</td>
<td>395</td>
<td>224</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>61-70</td>
<td>161</td>
<td>86</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>71+</td>
<td>46</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>75</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Ethnicity

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>White/European</th>
<th>Latin/Hispanic</th>
<th>Black/African/Caribbean</th>
<th>South Asian</th>
<th>East Asian</th>
<th>Middle Eastern</th>
<th>Mixed</th>
<th>Other</th>
<th>Prefer not to say</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4633</td>
<td>1494</td>
<td>195</td>
<td>72</td>
<td>52</td>
<td>102</td>
<td>178</td>
<td>339</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>102</td>
<td>339</td>
<td>0</td>
<td>72</td>
<td>178</td>
<td>41</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>178</td>
<td>140</td>
<td>5</td>
<td>156</td>
<td>399</td>
<td>42</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>156</td>
<td>399</td>
<td>41</td>
<td>5</td>
<td>175</td>
<td>90</td>
<td>42</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>90</td>
<td>42</td>
<td>6</td>
<td>25</td>
<td>7</td>
<td>13</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>7</td>
<td>13</td>
<td>2</td>
<td>136</td>
<td>71</td>
<td>37</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>136</td>
<td>71</td>
<td>37</td>
<td>11</td>
<td>147</td>
<td>41</td>
<td>27</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>147</td>
<td>41</td>
<td>27</td>
<td>39</td>
<td>99</td>
<td>23</td>
<td>37</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

**Materials**

It is widely documented that variability in encoding and test conditions is crucial when trying to detect reliable and generalizable effects (Brewer, Keast, & Sauer, 2010; Lindsay, Read, & Sharma, 1998). Accordingly, we created four 30 s, non-violent videos depicting four different crimes, so that encoding conditions varied on several dimensions, including [a] the appearance of the target (each video featured a different, white, male culprit); [b] the distinctive feature on the target (each culprit donned a unique distinctive feature); [c] the crime committed (carjacking, graffiti attack, mugging, theft), and [d] the exposure duration of the target in each video (which ranged from 5 to 16 s across the four videos). At test, variation occurred between the encoding stimuli (the target in the crime video) and the test stimuli (the target’s photographic image), simply because videos and photographs of people can vary to different extents. Targets also
varied in their similarity to the foils. A more complete description of each crime appears in the Supplemental Materials available online.

**Lineups**

We used 6-person simultaneous lineups that either contained the culprit and five foils (a “target-present” lineup), or contained six foils (a “target-absent” lineup). We created a pool of 40 foils for each culprit, so that we could randomly generate lineups from these pools. To create the pools of foils, we first asked a group of 18 subjects to watch each crime video and then answer 16 questions about the culprit’s physical attributes, including questions about gender, eye color, hair color, height, weight and ethnicity. Some characteristics required a categorical option choice (such as gender) whereas others required free-text responses (for instance, height and weight). In line with other studies (Carlson, Gronlund, & Clark, 2008; Zarkadi, Wade, & Stewart, 2009), we then entered the modal descriptions into the Florida Department of Corrections Inmate Database (http://www.dc.state.fl.us/AppCommon/) to retrieve 40 photographs of men who matched the modal description of each of the four culprits (160 total). This approach fits with the widely-accepted recommendation that foils should match the witness’s description of the culprit (Technical Working Group for Eyewitness Evidence, 1999; Wells, 1993).

The photos we selected from the database depicted men facing directly towards the camera. To control for the influence of emotional display, we selected men with neutral facial expressions (Flowe et al., 2014). We used Adobe Photoshop–CS5® to transform the images to greyscale and to remove any background color or pattern. If the person had a distinctive facial feature, we removed it. To prevent biases attributable to clothing, we also digitally altered each photo so that all foils appeared to be wearing a plain black t-shirt (Lindsay, Wallbridge, & Drennan, 1987). We took similar-looking “mug shots” of the culprits on the day we filmed the mock-crimes. We
edited these mug shots in the same way as the foil photographs, including adjusting the resolution to match that of the foil photographs.

Next, we edited the four pools of 40 (160 total) images to create foils for the replication, pixelation, and block lineups (see Fig. 1). For the replication lineups, we digitally added the culprit’s distinctive feature to each foil in the pool of 40. To reflect current police practice in several jurisdictions including England, Wales, New Zealand, Canada, and Germany, this distinctive feature was very similar in size, appearance and location—but not identical to—the culprit’s distinctive feature. For pixelation lineups, we concealed the culprit’s distinctive feature by pixelating it, and pixelating the same region on each of the 40 foils in the corresponding pool. For block lineups, we concealed the culprit’s distinctive feature by overlaying a solid black rectangle and we overlaid the same shape, in the same region, on each of the 40 foils in the corresponding pool. For target-present do-nothing lineups, we left the culprit’s distinctive feature uncovered and did nothing to the photos of the foils. In target-absent do-nothing lineups, we needed one foil face that had a distinctive feature similar to the culprit’s; accordingly, we used the replication foils, which had the culprit’s distinctive feature added. The other 5 foil photos in each target-absent lineup remained undoctored. Note that the target-absent do-nothing lineups mirror the real-world situation in which a witness reports the culprit’s distinctive feature to the police, but the police apprehend an innocent person with a similar distinctive feature and place him in the lineup.

To check that we had doctored our foils the way police actually doctor foils, we gathered evidence of ecological validity by consulting with a Detective Inspector from a local police force in the UK who sat on the National Committee for Identification Evidence. We randomly selected 18 foils to whom we had applied the replication, pixelation and block manipulation, and asked
her to evaluate them. The officer agreed that the images were concordant with police practice in England and Wales.

To make sure our replication foils did not look doctored, we then asked five new subjects to view all four replication foil pools (160 photos) and to identify any images that either did not match the modal description of the culprit, or looked as though they had been digitally altered. These subjects said that all the foils matched the descriptions of the culprits, but identified a total of 14 photos as looking as though they had been digitally altered. We then re-edited the distinctive features on these 14 photos until all five subjects were satisfied. Next, we asked a new group of 39 subjects to evaluate four target-present replication lineups (one for each culprit), in which the foils were randomly generated. We asked them to identify which photograph had not been digitally altered; they were no better than chance at this task (all ps > .20). Taken together, these findings suggest that our replication photos did not look manipulated, and our procedure for generating lineups did not bias subjects towards or against the suspect.

**Procedure**

Subjects were told that the study was about “Personality and Perception.” They were randomly assigned into one of the eight experimental conditions and one of the four crime videos (with the constraint that subject numbers were relatively equal in each condition).

There were three phases in the experiment. In the first phase, subjects watched a video of a crime. They were instructed to pay close attention because they would be asked questions about it later. After the video ended, we asked subjects if they had encountered any technical problems while viewing the video. The second phase, a filler phase, then began. In this phase, subjects worked on three questionnaires and an anagram puzzle for a total of 8 minutes. The questionnaires were the Autism Spectrum Quotient (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), the Six-Item Short-Form State scale of the Spielberger State-Trait
Anxiety Inventory (Marteau & Bekker, 1992), and the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003). We do not discuss subjects’ performance on these scales because they served as a filler task. In the third phase, we asked subjects to indicate their confidence that they would be able to recognize the culprit. Subjects responded on a 100-point Likert-type scale ranging from 1 (Completely Uncertain) to 100 (Completely Certain). Immediately after this task, subjects saw a lineup comprised of a $2 \times 3$ array of photos. Target-present lineups featured the culprit and five randomly selected foils from the corresponding pool. The position of the culprit was randomly determined for each subject. Replication, pixelation and block target-absent lineups consisted of six randomly selected foils (i.e. there was no designated innocent suspect). In do-nothing target-absent lineups, one foil with the culprit’s distinctive feature and five foils without the culprit’s distinctive feature were randomly selected (i.e. the innocent suspect was the foil that had the culprit’s distinctive feature). The position of the innocent suspect was also randomly determined for each subject. We chose this method of generating lineups to increase the generalizability of our results and to avoid the problems associated with using a small number of culprit - innocent suspect pairs. By randomly generating lineups, we also avoided using lineup fairness and bias measures, which are not always stable (Lindsay, Beaudry, Mansour, Bertrand & Kalmet, 2011).

All subjects were instructed that the culprit “may or may not be present” and then were asked to make a single identification by either clicking on the person they believed to be the culprit, or on an option labelled “Not Present.” Next, subjects used a 100-point Likert-type scale (1=Completely Uncertain to 100=Completely Certain) to rate their confidence in their decision. Finally, subjects answered a question that enabled us to check that they were paying attention (“What happened in the video that you watched?”), and they also answered a number of demographic questions.
Results

Recall that our primary aim was to determine the extent to which unfair lineups affect witnesses’ [a] willingness to identify the suspect, and [b] their ability to distinguish between real culprits and innocent suspects. We addressed these questions by using Receiver Operating Characteristic (ROC) analysis, and gathered further information by examining the distribution of subjects’ identification responses and subjects’ ability to judge the accuracy of their identification decisions.

ROC Analysis

Because the ROC approach is relatively new in the field of eyewitness memory, a brief overview should prove helpful. In ROC analysis, the first step is to construct an ROC curve for each lineup technique. Each curve plots the correct identification rate of guilty suspects in target-present lineups (hit rate; HR) against the false identification rate of innocent suspects in target-absent lineups (false alarm rate; FAR). In many ways, ROC analysis is like the traditional diagnosticity ratio, determined by HR/FAR (Steblay, Dysart, & Wells, 2011). But instead of calculating a single diagnosticity ratio (one HR/FAR pair), we plot several HR/FAR pairs over decreasing levels of confidence. Confidence serves as a proxy for willingness to choose, with decreasing levels of confidence equating to more liberal responding (Wixted & Mickes, 2014). Therefore, by plotting these HR/FAR pairs over the full range of confidence, we can determine how the different lineup types affect subjects’ ability to distinguish between real culprits and innocent suspects, independently of their willingness to identify the suspect (Gronlund, Wixted, & Mickes, 2014; National Research Council, 2014).

Figure 2 displays this thinking more concretely, and depicts two hypothetical ROC curves. The lowest left point of each curve, highlighted in grey, represents the correct and false identifications made at the highest level of confidence (“100% certain”). The second point on...
each curve represents the correct and false identifications made at the highest level of confidence and the second highest level of confidence (i.e. “100% certain”-“90% certain”), and so forth. As one moves along the curve, one eventually reaches the farthest right point (circled in grey), which represents all of the subjects in that condition who identified the suspect. A key idea is that for any point on the lower ROC (white circles), there is an achievable point on the higher ROC (solid black circles) that is associated with both a higher hit rate and a lower false alarm rate. Therefore, the ROC curve that falls closest to the upper left corner of the plot—closest to the star and farthest from the dashed chance line—is the objectively superior procedure because it maximizes culprit identifications while minimizing innocent suspect identifications. Put simply, this procedure allows witnesses to most accurately discriminate between culprits and innocent suspects.

To be clear, ROC analysis measures people’s ability to discriminate between guilty and innocent suspects, setting aside choices of known-to-be innocent foils. This is different to an absolute notion of memory discriminability—which would be the ability to discriminate between guilty suspects and anyone else in the lineup (i.e. innocent suspects and foils; see Wixted & Mickes, 2015, for a discussion). From a practical standpoint, discriminating between guilty and innocent suspects is arguably the key discriminability to measure because false identifications of foils do not result in any legal action against the foil that is selected. Nevertheless, we direct interested readers to our signal-detection modeling available in the Supplemental Materials online, because the modeling accounts for foil choices.

To compare ROC curves, we compare the partial Area Under the Curve (pAUC) because the false identification rate of innocent suspects is less than 1. In pAUC analysis, one defines the specificity (1 – FAR) for calculating the AUC. For example, if we were interested in the calculating the shaded area under the curve with the solid black circles in Figure 2, we would
calculate the $pAUC$ statistics by defining the specificity as $(1 - .09) = .91$. To calculate $pAUC$, we used the statistical package $pROC$, which also calculates a measure of effect size, $D$, using the formula: $D = (AUC_1 - AUC_2)/s$. In this formula, $s$ is the standard error of the difference between the two AUCs and is estimated using bootstrapping (Robin et al., 2011).

**Fig. 2.**

Two Hypothetical Receiver Operating Characteristic (ROC) Curves. The curve through the black operating points is the lineup procedure that allows witnesses to most accurately distinguish between real culprits and innocent suspects, because it falls closest to the ideal (the star) and furthest from the dashed chance line compared to the alternative procedure. The lowest left point on each curve (highlighted in grey) represents the correct and incorrect suspect identifications made with the highest level of confidence, whereas the farthest right
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

point (circled in grey) represents all of the subjects in that condition who identified the suspect. The partial Area Under the Curve (pAUC) for the shaded area under the curve with the solid black circles is calculated by setting the specificity (1 – FAR) to .91.

We now return to our empirical data. To construct our ROC curves, we collapsed the data across the four crime videos. We rounded subjects’ confidence ratings (made on a 100-point Likert scale) to the nearest 10 so that each curve would have 11 operating points of decreasing confidence (i.e. 100%, 90%, 80% and so forth). We then calculated the correct identification rates (HR) and the false identification rates (FAR) over the decreasing confidence levels. Correct identification rates were the number of guilty suspect IDs ÷ number of target-present lineups. The false identification rates were the number of innocent suspect IDs ÷ number of target-absent lineups.

We calculated innocent suspect identifications differently for the unfair and fair lineups. In the unfair (do-nothing) lineups, subjects made innocent suspect identifications when they identified the single lineup member with the distinctive feature. In the fair (replication, pixelation and block) lineups, recall that there was no designated innocent suspect—thus we estimated the number of innocent suspect identifications in these conditions using a common approach, dividing the number of false identifications made in target-absent lineups by the total number of people in the lineup—here, six (Brewer & Wells, 2006; Mickes, 2015). The procedure works on the assumption that the lineup member that best matches the subject’s memory of the culprit is the innocent suspect (Palmer, Brewer, Weber, & Nagesh, 2013). One particular benefit of estimating false identifications in this way is that it leads to a more conservative measure of false identifications. Because the innocent suspect may not always be the most similar in appearance to the actual culprit, this method of estimation can only overestimate, not underestimate, the
number of false identifications in target-absent lineups. Thus, using this estimation method in replication, pixelation and block lineups provided a conservative test of how well these (fair) techniques enhance witness identification performance compared to the (unfair) do-nothing lineups.

Figure 3 shows the ROC curves for the fair and unfair lineups. When calculating pAUC statistics, we set the specificity to 0.91—which corresponded to the FAR range covered by the least extensive curve (block; FAR range: 0 to .09)—for two main reasons. First, by setting the
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

FAR range from 0 to .09, we prevented the pROC program from having to extrapolate the three fair lineup curves over a vast range (from a FAR of .09 to .40). The pROC program uses a crude method of extrapolation, so doing so over large distances can reduce statistical accuracy. Second, the lower FAR range (0 to .09) may have greater practical relevance, because the legal system [a] is interested in knowing which conditions increase witnesses’ ability to distinguish between innocent and guilty suspects when the false alarm rate is low, and [b] may take these high-confidence identifications more seriously than low-confidence identifications (see Gronlund et al., 2012). We are confident that limiting the pAUC analysis to a small subset of the do-nothing curve did not affect our findings. When we fit a signal-detection process model of lineup performance to our data we found the same pattern of results (see online Supplemental Materials). This modeling technique uses the largest FAR range that a target-absent lineup can support.

To what extent did our lineup types affect witnesses’ performance? More specifically, did the unfair lineups increase witnesses’ willingness to choose the suspect—or did those lineups impair witnesses’ ability to distinguish between the culprit and the innocent suspect? As Figure 3 shows: Compared to the replication, pixelation and block (i.e., fair) lineup techniques, doing nothing increased subjects’ willingness to identify the suspect and also markedly impaired subjects’ ability to discriminate between real culprits and innocent suspects. Focusing on the ROC curves in Figure 3, we can see that the do-nothing ROC points have shifted more to the right than any of the fair lineup ROC points. This shift right shows there was an increase in both correct and false identifications. That is, subjects’ willingness to identify the suspect increased in the do-nothing lineups, as compared to replication, pixelation and block lineups.

A more striking finding though, is that do-nothing lineups made it more difficult for subjects to distinguish between innocent and guilty suspects. The pAUC for do-nothing lineups (pAUC =
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

0.008, 95% CI: 0.006, 0.010) was significantly smaller than the pAUC for replication (pAUC = 0.016, 95% CI: 0.013, 0.019, \( D = 4.11, p < .001 \)), block (pAUC = 0.016, 95% CI: 0.013, 0.019, \( D = 4.35, p < .001 \)) and pixelation (pAUC = 0.015, 95% CI: 0.012, 0.018, \( D = 4.17, p < .001 \)) lineups. Finally, the three fair lineups led to similar levels of identification performance—the pAUCs did not differ significantly between replication and block (\( D = 0.08, p > .250 \)), replication and pixelation (\( D = 0.32, p > .250 \)), or block and pixelation (\( D = 0.24, p > .250 \)) lineups. We also fit a signal-detection process model of lineup performance to our data to further confirm these findings (see Lampinen, in press; Wixted & Mickes, 2014). The modeling procedure and results are presented in the Supplemental Materials available online. Importantly, the model fitting exercise and our pAUC analysis led to the same results. Taken together, these findings fit with the additional prediction of the diagnostic-feature-detection model that doing nothing to stop distinctive suspects from standing out does not just increase witnesses’ willingness to choose the suspect, it also markedly impairs their ability to sort guilty and innocent suspects into their appropriate categories.

Identification Responses

To further understand the effect of unfair lineups on subjects’ identification performance, we calculated the proportion of suspect identifications, foil identifications and lineup rejections for each lineup type. Table 2 shows the frequencies and percentages of identification responses for each lineup type. There is an interesting point to note about these data. We know from the ROC analysis that unfair lineups led to more suspect identifications than fair lineups. The data in Table 2 indicate that this overall increase in suspect identifications was accompanied by a decrease in both foil identifications and in lineup rejections in target-present lineups, but just a decrease in foil identifications in target-absent lineups.
Table 2.

*Frequencies and Percentages of Identification Responses in Replication, Pixelation, Block and Do-nothing Lineups.*

<table>
<thead>
<tr>
<th>Identification Responses</th>
<th>Replication</th>
<th>Pixelation</th>
<th>Block</th>
<th>Do-nothing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Target-present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suspect</td>
<td>347.00</td>
<td>30.84</td>
<td>320.00</td>
<td>27.95</td>
</tr>
<tr>
<td>Foil</td>
<td>382.00</td>
<td>33.96</td>
<td>411.00</td>
<td>35.90</td>
</tr>
<tr>
<td>Reject</td>
<td>396.00</td>
<td>35.20</td>
<td>414.00</td>
<td>36.16</td>
</tr>
<tr>
<td>Target-absent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suspect</td>
<td>104.50</td>
<td>9.17</td>
<td>102.33</td>
<td>9.10</td>
</tr>
<tr>
<td>Foil</td>
<td>522.50</td>
<td>45.83</td>
<td>511.67</td>
<td>45.52</td>
</tr>
<tr>
<td>Reject</td>
<td>513.00</td>
<td>45.00</td>
<td>510.00</td>
<td>45.37</td>
</tr>
</tbody>
</table>

*Note.* Collapsed over all four mock-crime videos. In target-absent do-nothing lineups, suspect identifications are identifications of the lineup member with the distinctive feature, whereas foil identifications are identifications of any other lineup member without the distinctive feature. In replication, pixelation and block target-absent lineups there was no designated innocent suspect. We estimated the number of innocent suspect identifications by dividing the total number of false identifications in target-absent lineups by six (the number of faces in the lineup). Similarly, we estimated the number of foil identifications by dividing the total number of false identifications by six (the number of faces in the lineup), and then multiplying by five (the number of faces that were not the innocent suspect).

Focusing on target-present lineups first: A 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (ID response: suspect, foil, incorrect rejection) chi-square analysis showed that
lineup type influenced ID responses, $\chi^2 (6, N = 4507) = 282.70, p < .001, V = .18$. Specifically, fair lineups led to fewer suspect IDs (replication: $z = -2.84, p < .01$; pixelation: $z = -4.50, p < .001$; block: $z = -4.07, p < .001$) but more foil IDs, (replication: $z = 1.90, p > .05$; pixelation: $z = 3.09, p < .01$; block: $z = 2.29, p < .05$) and more rejections (replication: $z = 1.13, p > .05$; pixelation: $z = 1.70, p > .05$; block: $z = 2.02, p < .05$), than expected. Conversely, unfair lineups led to more suspect IDs ($z = 11.53, p < .001$), but fewer foil IDs ($z = -7.36, p < .001$), and fewer rejections ($z = -4.90, p < .001$), than expected. In short, when the suspect was left to stand out in target-present lineups, there was an increase in suspect identifications along with a reduction in both foil identifications and incorrect rejections.

Next, focusing on target-absent lineups, recall that in replication, pixelation and block target-absent lineups there was no designated innocent suspect. We therefore estimated the number of innocent suspect identifications by dividing the total number of false identifications by six (the number of faces in the lineup). Similarly, we estimated the number of foil identifications by dividing the total number of false identifications by six (the number of faces in the lineup), and then multiplying by five (the number of faces that were not the innocent suspect). A 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (ID Response: suspect, foil, correct rejection) chi-square analysis using these estimates showed that lineup technique influenced ID responses, $\chi^2 (6, N = 4418) = 481.70, p < .001, V = .23$. Fair lineups led to fewer innocent suspect identifications (replication: $z = -5.22, p < .001$; pixelation: $z = -5.24, p < .001$; block: $z = -5.50, p < .001$) but significantly more other foil identifications (replication: $z = 3.26, p < .001$; pixelation: $z = 3.08, p < .001$; block: $z = 2.38, p < .001$), than expected. Conversely, unfair lineups led to more innocent suspect identifications ($z = 16.85, p < .001$), but significantly fewer other foil identifications ($z = -9.21, p < .001$), than expected. The proportion of correct rejections in all four lineup types was similar (replication: $z = -0.03, p > .05$; pixelation: $z = 0.15, p > .05$;
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

block: $z = 0.954$, $p > .05$; do-nothing: $z = -1.14$, $p > .05$). This analysis indicates that when the suspect was left to stand out in target-absent lineups, subjects shifted their identifications from the other lineup members, onto the innocent suspect.

**Confidence and Accuracy**

Recall that the diagnostic-feature-detection model suggests that unfair lineups impair a witness’s ability to distinguish between innocent and guilty suspects because it is not obvious to the witness that the suspect’s distinctive feature is unhelpful. If witnesses fail to realize that the distinctive feature is unhelpful, they may not lower their confidence judgment to compensate for their poorer performance. If this account is correct, then subjects who viewed the unfair do-nothing lineups should be less accurate, at every level of confidence, than subjects who viewed the fair replication, pixelation and block lineups.

To test this prediction, we plotted suspect identification accuracy (correct IDs of guilty suspects in target-present lineups ÷ correct IDs of guilty suspects in target-present lineups + false IDs of innocent suspects in target-absent lineups) separately for each level of confidence (100%, 90%, 80% and so forth, as per Mickes, 2015). This method of calculating suspect identification accuracy reflects the probability of guilt, given that the suspect was identified (i.e. the posterior probability of guilt). We estimated the number of innocent suspect identifications in the replication, block and pixelation lineups in the same way we did for the ROC analysis. To provide more stable estimates, confidence level was binned into five categories (0–20%, 30–40%, 50–60%, 70–80%, 90–100%, see Brewer & Wells, 2006). For interested readers, the frequencies of identification responses in each confidence bin in replication, pixelation, block and do-nothing lineups are shown in Table S1 in the Supplemental Materials available online.
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

Fig. 4.
Confidence-Accuracy Curves for the Fair (Replication, Pixelation, Block) and Unfair (Do-nothing) Lineups. Bars represent standard error bars. The dashed diagonal line depicts chance accuracy at the lowest confidence bin and perfect accuracy the highest confidence bin.

Figure 4 shows the confidence-accuracy curves for each lineup type. Non-overlapping error bars denote reliable differences between the lineup techniques (e.g. Sauer, Brewer, Zweck, & Weber, 2010). As predicted, subjects who viewed the unfair, do-nothing lineups showed lower levels of accuracy at every level of confidence than subjects who viewed the fair lineups. Put another way, an identification made at any level of confidence from an unfair lineup was less trustworthy than an identification made with the same level of confidence from a fair lineup.
These data align with the diagnostic-feature-detection model, which suggests that when nothing was done to stop the distinctive suspect from standing out, subjects may have been unaware that their memory accuracy was worse and therefore failed to adjust their confidence accordingly.

**Discussion**

In this paper, we asked why unfair lineups promote mistaken identifications. Our findings suggest that unfair lineups—in comparison to fair lineups—make people more likely to identify the suspect, but worse still, unfair lineups impair people’s ability to distinguish between guilty and innocent suspects and distort people’s ability to judge the trustworthiness of their identification decision.

It is arguably unsurprising that our unfair lineups, in which a suspect was left to stand out, increased subjects’ willingness to identify that suspect. Many eyewitness identification studies have demonstrated this already (Clark, 2012; Doob & Kirshenbaum, 1973; Fitzgerald et al., 2013; Wells et al., 1979; Wells et al., 1993). The fascinating finding is that unfair lineups also dramatically hindered subjects’ ability to sort innocent and guilty suspects into their appropriate categories. This mechanism has not been discussed until now, yet it is important. Procedures that simply make witnesses less willing to choose the suspect decrease innocent suspect identifications but also come at a cost: they stifle culprit identifications (Clark, 2012). Procedures that enhance a witness’s ability to distinguish between innocent and guilty suspects minimize innocent suspect identifications and maximize culprit identifications, regardless of the witness’s willingness to choose. Arguably then, this is the critical mechanism to investigate (Gronlund, Wixted, & Mickes, 2014; National Research Council, 2014).

So, why might unfair lineups harm people’s ability to distinguish between the real culprit and an innocent suspect? One explanation is that witnesses fail to appreciate that the suspect’s distinctive feature is not useful in an unfair lineup and so rely heavily on it to make their
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

identification. By contrast, when lineups are fair and the suspect does not stand out, witnesses can appropriately discount the distinctive feature and give more weight to other, more informative, cues (Wixted & Mickes, 2014). Support for this theoretical account comes from the finding that, in the unfair lineups, subjects failed to compensate by setting a more conservative confidence criterion when making an identification. This fits with a mechanism in which subjects do not realize that their accuracy is impaired.

Importantly, a growing body of research suggests that mock-witnesses are generally good at judging the likely accuracy of their memories even when their accuracy is impaired (e.g. Brewer & Wells, 2006; Mickes, 2015, Exp.1; Palmer et al., 2013; Sauer et al., 2010). Palmer et al., for instance, showed that divided attention significantly impaired people’s memory ability, yet, when the authors plotted accuracy at each level of confidence it didn’t matter if subjects had full or divided attention at encoding—subjects’ accuracy at each level of confidence was generally the same (Exp.2, see Figures 3 & 4). Palmer and colleagues concluded that their experimental manipulations did not undermine the usefulness of confidence as an indicator of accuracy. This study, and many others, shows that people typically recognize when their memories are poor and adjust their confidence appropriately (Mickes, 2015, Exp.1; Sauer et al., 2010; Palmer et al. 2013, Exp.1). There are, however, some instances in which confidence is uninformative of accuracy (e.g. Chandler, 1994; Mickes, 2015, Exp.2). Indeed, our findings show that unfair lineups can systematically distort confidence.

One consequence of identifications from unfair lineups being less accurate at every level of confidence, is that subjects in the do-nothing condition made high-confidence suspect identifications (90-100% certain) when accuracy was moderate (60%). This finding has serious implications for criminal justice because legal decision makers are strongly influenced by highly confident witnesses (Brewer & Burke, 2002; Wells, Lindsay, & Ferguson, 1979). Although
subjects in the fair lineup conditions (i.e., replication, pixelation or block) were under-confident at the lower end of the confidence scale, the critical point is that their identifications were consistently and substantially more trustworthy. Moreover, subjects who viewed the fair lineups only identified the suspect with high confidence (90-100% certain) when they were very likely to be accurate (> 80% accurate). Therefore, highly-confident suspect identifications made from replication, pixelation and block lineups are likely to be very informative for triers of fact (also see Supplemental Materials for further discussion on subjects’ confidence ratings).

At first glance, our results appear to conflict with two face-recognition studies that suggest replicating distinctive features is better than removing them (Badham, Wade, Watts, Woods, & Maylor, 2013; Zarkadi et al., 2009). Zarkadi and colleagues, for example, found that replication increased correct identifications by approximately 20% in target-present lineups, while we found replication and concealment techniques were equally effective. There is, however, a crucial methodological difference to consider. The previous research compared replication lineups with removal lineups in which the target’s distinctive feature was simply removed. Subjects made more incorrect rejections in target-present removal lineups possibly because the person they believed to be the culprit was now missing a prominent distinctive feature that they remembered (Wixted & Mickes, 2014). Subjects in our study were unlikely to use this strategy because we tested pixelation and block lineups, both of which indicate that there could be a distinctive feature underneath the concealed area. Therefore, unlike the previous research, we did not observe a relatively high number of incorrect rejections in pixelation and block lineups compared to replication lineups. Instead, we observed similar performance in all three fair conditions.

On a practical level, our research suggests that law enforcement officers should take steps to prevent distinctive suspects from standing out. If unfair lineups just increased witnesses’ willingness to choose the suspect (and did not affect their ability to distinguish between innocent
UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN

and guilty suspects), then officers could remedy this by inducing more conservative responding. For instance, urging witnesses to be cautious (“Be certain before making a decision”) should increase the amount of memorial information witnesses demand before choosing and result in fewer positive, and therefore fewer suspect, identifications (Clark, 2005). Our data, however, suggest that law enforcement officers need to apply fair lineup techniques to improve identification accuracy, and that replication, pixelation or block techniques are equally effective.

In sum, our data fit the predictions of a new model, the diagnostic-feature-detection model. Testing theoretical models is important, because, once refined, theories can be used to develop procedures that further enhance eyewitness accuracy. More specifically, our findings shed light on the processes underlying the harmful effects of unfair lineups and suggest that when suspects are unduly distinctive, witnesses are not just more willing to choose the suspect, they also struggle to distinguish between guilty and innocent suspects. Perhaps if Leonard Callace had been placed in a fair lineup, alongside foils who also had full beards or whose chins had been concealed, he would not have spent six years in prison for a crime he did not commit.

References


UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN


UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN


UNFAIR LINEUPS DON’T JUST MAKE WITNESSES CHOOSE MORE OFTEN


**Author Contributions**

K. A. Wade and M. F. Colloff developed the study concept and design. Data collection were performed by all authors and M. F. Colloff performed the data analysis and interpretation under the supervision of K. A. Wade. M. F. Colloff and K. A. Wade drafted the manuscript, and D.
Strange provided critical revisions. All authors approved the final version of the manuscript for submission.

**Acknowledgements**

We thank John Wixted and Laura Mickes for their insightful discussions and indispensable practical advice. We also thank Neil Stewart for valuable programming assistance, and Maryanne Garry and Elizabeth Maylor for their helpful comments on an earlier draft.